Faculty of Science and Engineering
School of Electrical Engineering, Computing and Mathematical Sciences

Acoustic Speaker Localization with Strong Reverberation and Adaptive Feature Filtering with a Bayes RFS Framework

Shoufeng Lin

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

February 2019

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Shoufeng Lin, 2019/02/22

Acknowledgements

Pursuing this PhD has been a life changing experience in the past six years. Out of pure interest, I found it most enjoyable in exploring those challenging problems in the speech signal processing area. Perhaps the most challenging part however is the times of ultimate helplessness, especially when everything around seemed to have gone astray. It takes much more than the continuous efforts and courage. I am much humbled.

There are many people whom I have had the opportunity to come across in this journey and would like to express my sincere gratitude to.

First of all, I would like to thank my supervisor Prof. Sven Nordholm. Thank you for leading me into the study of speech signal processing, your special strictness and bearing of the pains, and your dedicated efforts in teaching me humbleness during my exploration.

I also thank Prof. Ba-Tuong Vo who helped me to get started with the learning of the Bayes RFS tracking filters. Thank you for your kind advice.

My gratitude also goes to my thesis committee chair Prof. Yue Rong. Thank you for all your precious time and suggestions at my difficult times.

I owe my thanks to my dear friend Dr. Tze-Chuen Toh. Thank you for your willingness and help to review my writings even though you are super busy!! Your kindness and encouragement lighted the dark road. I also would like to thank Prof. Jont Allen, Prof. Haizhou Li, Prof. Dan Ellis, Prof. Ian Howard, Dr. Yee-hong Leung, Dr. Grace Yun, Dr. Ahmed Abu-Siada, Dr. Erfan Loweimi, Dr. CC Lai, Dr. Linh Tran, Dr. Renato Nakagawa, and many more. Thank you for the kind sharing, advice, good wills and being there. I wish that we could do some interesting things together in the near future.

The Australian Postgraduate Award, Australian Government Research Training Program scholarship and my engineering and teaching jobs helped me to support my family and kept my composure during the earlier stage.

Finally and foremost, I am in debt to my parents Shifa Lin and Meirong Lin for their unconditional love and support, and my loving wife Dr. Ting Lin and lovely kids. There is something that can never change...

Abstract

This thesis presents the investigation of several challenging parameter and state estimation problems in the signal processing paradigm, viz. the speaker localization in presence of reverberation, multi-speaker tracking, and multi-feature multi-speaker state filtering, using microphone recordings.

Acoustic speaker localization has been a long-standing signal processing challenge, especially in reverberant environments and when speakers are moving. Thus the first part of the thesis focuses on reliably estimating speaker locations in short time intervals, and three algorithms are developed for the reverberation-robust localization, respectively in the time and the frequency domains. The first two algorithms are built upon the voiced speech signal and room impulse response (RIR) models. A novel onset detection and encoding scheme is derived to prefilter the direct-path cues, which are then used to formulate the cross-correlation coefficients for reliable localization. The third algorithm is built on the classic generalized cross-correlation - phase transform (GCC-PHAT) method and a room transfer function (RTF) model. It exploits the redundant information from multiple microphone pairs to suppress the effect of sound reflections. Performance evaluation in various reverberant conditions demonstrates the benefits of the proposed localization algorithms compared with the state-of-the-art methods.

Multi-speaker tracking has also captured increasing attentions from the research communities in the past two decades. Estimating speaker states with correct identities has been one of the main challenges, especially when the number of speakers is unknown and time-varying. Thus the second part of the thesis explores the adaptive speaker feature filtering, where the location estimates from the first part are treated as observations of random speaker kinematic states. The state-of-the-art generalized labeled multi-Bernoulli (GLMB) Bayes random finite set (RFS) filter is used as the basis of the proposed speaker feature state filtering framework. The measurement-driven birth (MDB) model for the GLMB filter is implemented for adaptive filtering. Two typical scenarios of practical importance are investigated. The first one estimates the kinematics feature state only and produces labeled

trajectories of respective speakers. Performance of the proposed framework is demonstrated in comparison to the well-known cardinalized probability hypothesis density (CPHD) filter. The second scenario further investigates the feasibility of generalizing the state filtering of a single feature into that of multiple features. The location, pitch and sound of each speaker are accommodated as a state vector, incorporating the independently transitioning and non-transitioning features. Experimental results show that the proposed multi-feature multi-speaker state filtering framework can jointly track and separate locations, pitches and sound signals of multiple speakers.

Concluding remarks, interesting future works, appendices and the bibliography are provided in the third part of the thesis.

Keywords: Speaker localization, reverberation, moving speakers, RIR, RTF, speech onset, redundant information, adaptive multi-speaker tracking, adaptive multi-feature multi-speaker filtering.

List of Acronyms

ASR	automatic speech recognition
CASA	computational auditory scene analysis
CPHD	cardinalized probability hypothesis density
CRB	Cramér-Rao bound
CSD	cross-power spectral density
DOA	direction of arrival
FISST	finite set statistics
GCC	generalized cross-correlation
GLMB	generalized labeled multi-Bernoulli
i.i.d.	independent and identically distributed
MCCC	multi-channel cross-correlation coefficient
MAP	maximum *a posteriori*
MDB	measurement driven birth
MLE	maximum likelihood estimation / estimator
NLP	natural language processing
PDF	probability density function
PHAT	phase transform
RFS	random finite set
RIR	room impulse response
RMSE	root-mean-square error
SMC	sequential Monte Carlo
SNR	signal to noise ratio
SSS	strong (strict) sense stationary
STFT	short time Fourier transform
TDOA	time difference of arrival
UCA	uniform circular array
WGN	white Gaussian noise
w.r.t.	with respect to
WSS	wide (weak) sense stationary

List of Symbols

Efforts have been made to represent parameters with unique symbols. Due to limited choices however, duplicated use of symbols may be occasioned but meanings are made clear in respective context.

α	constant weight for wideband beamformer
b	subband index
c	cut-off parameter of OSPA metric
d	index of zero crossing
$\mathbf{d}$	steering vector of microphone array
$\bar{d_p}^{(c)}$	OSPA metric
$\epsilon()$	localization function
$f_+(\cdot)$	state transition pdf
$g(\cdot)$	measurement probability function
h	impulse response
$\hat{\mathrm{h}}_{qi}$	the realization of impulse response from source q to microphone i
h_{qi}	the statistical model of the impulse response
i	sensor / microphone index
I_M	total number of sensors / microphones
j	sensor / microphone index
$\jmath$	$\sqrt{-1}$
$\mathcal{J}$	cost function
k	discrete signal frame index
ℓ	label of single object state
m	discrete signal sample index
$\overrightarrow{m_i}$	location of microphone i
n	discrete signal sample index
N	number of samples
p	order parameter of OSPA metric
P	beamformer response
$\wp$	source location
q	source index
r_a	radius of the circular microphone array

r_s	distance from source to sensor array
s	single-object state vector (without label)
$\mathbf{s}$	labeled single-object state
$\mathbf{S}$	labeled multi-object state
s_q	speech signal of source q in time domain
S_q	speech signal of source q in frequency domain
t	continuous time
τ	time delay
θ	DOA
$\boldsymbol{\theta}$	DOA vector
ϑ	label association
u	zero crossing index
v	velocity of sound
ν	weighting function for wideband beamformer
w	weight of window function / probability of hypothesis
$\mathbf{w}$	weight vector for beamformer
ω	angular frequency of digital signal (normalized) / probability of hypothesis
Ω	angular frequency of analog signal
x_i	signal recorded by microphone i
$\mathbf{x}$	vector of signals in time domain
$\mathbf{X}$	vector of signals in frequency domain
y	estimated signal
z	measurement
Z	set of measurement
π	multi-object probability density function
$\boldsymbol{\pi}$	GLMB multi-object probability density function
$\mathbb{R}$	the set of real numbers
$\Re$	the real part of a number
$\Im$	the imaginary part of a number
$*$	convolution operator
$\mathbb{E}$	mathematical expectation
$[\cdot]^T$	matrix transpose operator
$[\cdot]^H$	matrix conjugate transpose (Hermitian) operator

$[\cdot]^{-1}$	inverse of matrix
$[\cdot]^{+}$	Moore-Penrose pseudoinverse of matrix
$[\cdot]_{+}$	Bayesian recursion prediction
$\det[\cdot]$	matrix determinant
$\odot$	Hadamard product
$\otimes$	Kronecker tensor product

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

Since the turn of the 21st century, multi-speaker online tracking using microphone arrays has been an emerging research and engineering problem [1–6]. Such trend echoes with the increasing demand and applications of the automatic camera steering for online lecturing, remote conferencing and virtual reality [7,8], as well as speech separation and recognition front-ends [9,10]. In this problem, acoustic data recorded by microphones are mixtures of concurrent speech signals and their time-delayed reflections beside noise, where the number of active speakers is unknown *a priori* and time-varying, and each speaker can be moving and competing with others at any time instant.

Solving this problem involves addressing three major challenges in a broad sense, namely a) the multi-object state estimation (time-varying speaker states), b) the localization of moving sources of nonstationary wideband signals (human speech), and c) the multipath effect (acoustic reverberation). The desired outcomes are the separated trajectories of respective speakers. Here a trajectory can be regarded as time-indexed locations of an active speaker, and they are labeled so that each of the trajectories is uniquely associated with one of the speakers. Despite the amazing ease of human listeners in dealing with this problem, a perfect signal processing solution has not yet been found in the literature. The main focus of this thesis is thus an original systematic framework for localizing and tracking multiple moving speakers using microphones, with the attention to the challenge of reverberation.

The study of state estimation algorithms can date back to the early development of Kalman filter about half century ago [11]. Based on the linear and Gaussian dynamical and measurement statistical models, the Kalman filter keeps track of the system state over time via a prediction-update Bayesian recursion, supplied with measurement data. Besides the ubiquitous Kalman filter and its extensions for nonlinear systems, e.g. the extended Kalman

filter (EKF) [11] and the unscented Kalman filter (UKF) [12], the nearest neighbour (NN) [13], the multiple hypothesis tracking (MHT) [14] filter and the joint probabilistic data association (JPDA) [13] and their respective variants and extensions can also been found in the literature. Usually the NN scheme is susceptible to false alarms and miss-detections, the MHT requires an exhaustive search of measurement-to-object associations, while the JPDA assumes a known number of objects. In contrast, the Bayes RFS filters are the first close-form solutions to multi-object tracking [15–23], built upon the the finite set statistics (FISST) theory [24]. The Probability Hypothesis Density (PHD) [15, 16] and the cardinalized PHD (CPHD) [17, 19] have been well-accepted members in the family of multi-object Bayes RFS filters, and the CPHD has shown favourable performance over the PHD filter [19]. However, CPHD only propagates the first-moment (intensity) over the state space and cardinality distributions. The most current development of the Bayes RFS kind is the Generalized Labeled Multi-Bernoulli (GLMB) filter, which can track multi-object labeled states. Thus the GLMB Bayes RFS framework [21–23] is used for multi-speaker tracking, in order to systematically produce labeled state estimates of multiple speakers. To complement the system for practical applications, the measurement-driven object birth model for GLMB [25] is implemented for adaptive filtering, and due considerations for multi-sensor measurements are also necessary.

Sound source localization is a long-standing yet very challenging signal processing research topic. Location estimation algorithms constitute a critical part of the multi-speaker tracking framework, as they provide the measurement data for the state estimation. Numerous location estimators can be found in the literature, beginning from the 1960s [26]. Representative and influencing works include the subspace based methods, e.g. the multiple signal classification (MUSIC) algorithm and the estimation of signal parameters via rotational invariance techniques (ESPRIT), [27–30], steered response power (SRP) beamformers [31–34], and time delay estimation (TDE) or time difference of arrival (TDOA) based methods [35–39]. Most of these localization methods, are based on some simplifying assumptions, e.g. narrowband signals, static acoustic sources or anechoic environments,

and thus may suffer when these assumptions are violated in practice. Wideband extensions are available, e.g. the eigenbeam (EB)-ESPRIT [40] and the time-frequency circular harmonic beamformer (TF-CHB) [34]. However, they are restricted to uniform circular arrays (UCAs), and may still suffer from performance degradation in presence of reverberation and moving sources, due to the underlying assumptions. The multichannel cross-correlation coefficient (MCCC) method [41, 42] suggests using the redundant information from multiple microphones. The reverberation-robust SRP - phase transform (PHAT) [43, 44] extends the classic generalized cross correlation - phase transform (GCC-PHAT) method, but has a modest spatial resolution. A Neuro-Fuzzy [39] approach mimicking the human auditory functions is found applicable to moving speakers with reverberation, although it was not a result of rigorous derivations from signal models. Therefore, this thesis develops three robust localization algorithms, viz. the onset - generalized steered response power (Onset-GSRP), onset - multichannel cross-correlation (Onset-MCC) and MCC-PHAT [45], which are reliable for the localization of multiple moving speakers in presence of reverberation. In particular, the Onset-MCC and MCC-PHAT have good spatial resolutions. The localization results can have gaps or miss-detections of speaker locations due to the nonstationarity of speech signals and interference of concurrent speakers, and clutter or spurious estimates due to reverberation. Moreover, localization alone does not provide measurement-to-speaker association. Therefore, filtering of the location candidates via the multi-object state estimator is often needed.

1.2 Key Contributions

This thesis addresses the problem of speaker localization and tracking, with a focus on the challenges of the acoustic reverberation and moving speech sources, and presents a systematic implementation of the adaptive GLMB Bayes RFS approach in the multi-speaker state filtering. The main contributions are summarized as follows.

For the localization of moving speakers with reverberation:

- Two main approaches are proposed. The first approach (including two algorithms) builds upon the RIR model and develops a novel subband onset detection and encoding method for extracting direct-path cues. Although inspired by the computational auditory scene analysis (CASA) techniques [46] and psychoacoustic inferences, the proposed onset detection and encoding method is derived based on the speech signal and acoustic RIR models. Cross-correlation coefficients of the direct-path cues are then formulated and from relative sample delays into the spatial locations, for computationally efficient implementation. The resulting algorithms are referred to as the Onset-GSRP and the Onset-MCC, where the latter has improved spatial resolution.

- The second approach studies the acoustic RTF model and exploits the redundant information from multiple microphone pairs to suppress the effect of reverberation. Thus based on the classic GCC-PHAT method, the MCC-PHAT method is proposed with improved resolution and robustness against reverberation. Performance of all the proposed localization methods are compared with other state-of-the-art location estimators, using simulated sound signals and real-world recordings in various reverberant conditions. Improved spatial resolution and the robustness against reverberation and moving speakers are demonstrated.

For the multi-speaker tracking (feature filtering):

- An adaptive multi-speaker tracking filter is developed, based on the proposed localization methods and the GLMB Bayes RFS multi-object tracking framework. A measurement-driven birth model is used for adaptive online filtering. Performance of the proposed tracking system is evaluated and compared with the framework using the well-accepted CPHD filter. The presented adaptive GLMB filter provides a closed-form solution to the multi-speaker tracking, and jointly associates speaker kinematic states with respective identities (labels).

- A novel multi-feature multi-speaker tracking-and-separation filter is proposed, to verify the feasibility of generalizing the single feature state

filtering to that of multiple features, and to resolve the ambiguity in tracking a single feature state. The measurement-driven birth model is also used for the adaptive filtering. The proposed multi-feature multi-speaker state filter is shown applicable for jointly tracking and separating locations, pitches and sound signals of respective speakers.

1.3 Outline of the Thesis

Chapter 2 gives the background of the thesis, basic models for the speech signal, RIR, RTF and microphone array representations, formulates the localization and tracking problems, and reviews existing methods. Studies and analysis of the existing methods clarifies the need for reverberation-robust location estimators and an adaptive multi-object tracking framework.

The rest of the thesis is mainly composed of three parts.[1]

Part I concerns the location estimation. It focuses on the reverberation-robust localization especially for moving speakers. Specifically,

- Based on the speech signal and acoustic RIR models, Chapter 3 derives the novel onset detection and encoding algorithms, for the non-stationary voiced speech signals in presence of the multipath effect. The Onset-GSRP and Onset-MCC use the encoded speech onsets as direct-path cues, and formulate cross-correlation coefficients for reliable localization.

- The MCC-PHAT method builds upon the acoustic RTF model and the classic GCC-PHAT method, and exploits the redundant information from multiple microphone pairs, which is shown useful for reverberation robustness. In the numerical studies, the performance of the proposed speaker localization methods is demonstrated and compared with the state-of-the-art methods in various reverberant environments.

Part II investigates the multi-speaker feature filtering. Specifically,

[1] Due to the relative scope of the work, the thesis builds upon, and hence assumes that readers have reasonable understandings in, signals and systems, digital signal processing, probability and stochastic processes, parameter estimation and state filtering theories.

- Chapter 4 presents a systematic implementation of the GLMB Bayes RFS filter in adaptively tracking multi-speaker kinematic states, using the location estimates from the Onset-MCC and MCC-PHAT as proposed in Chapter 3. The GLMB filter is used and its performance is compared with that of the CPHD filter. Labeled kinematic states of multiple speakers are filtered, supplied with location estimates.

- In Chapter 5, a novel extension of the adaptive GLMB filter for tracking and separating multiple speaker features is also proposed. A proof-of-concept implementation is presented, which can jointly track and separate locations, pitches and sound signals of multiple speakers.

Part III concludes the thesis in Chapter 6, prospects some possible future works, and supplements in Appendices detailed derivations for the presented works.

1.4 Peer-reviewed Papers

The published papers arising from the current research are listed as follows:

- Reverberation-Robust Localization of Speakers Using Distinct Speech Onsets and Multichannel Cross-Correlations, by **Shoufeng Lin**; IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 2018 Nov 1;26(11):2098-111. [45]

  Some materials of this paper appear in Chapter 3. Permission for including this paper in the thesis can be found in Appendix L .

- Jointly Tracking and Separating Speech Sources Using Multiple Features and the generalized labeled multi-Bernoulli Framework, by **Shoufeng Lin**. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018). [6]

  The materials of this paper appear in Chapter 5. Permission for including this paper in the thesis can be found in Appendix L .

Chapter 2

Background and Literature Review

This chapter gives the background of the studied problem, provides the underlying signal and system models for problem formulation and solution derivation, and sketches an overview of existing state-of-the-art speaker localization and multi-object tracking algorithms. In particular, Section 2.2 and Section 2.3 provide the signal and system models in the time and frequency domains. Microphone array signals are modeled in Section 2.4. Section 2.5 and Section 2.6 offer an overview of the localization and tracking literature respectively. Finally, Section 2.7 formulates the speaker localization and tracking tasks as estimation problems.

2.1 Speaker Localization and Tracking

There is a significant body of literature on the problem of acoustic speaker localization and tracking. Obtaining accurate location estimates enables further signal processing, e.g. speaker tracking, speech separation and enhancement. It also has wide practical applications, such as the automatic camera steering in smart environments, finding sound source directions in hearing aids and smart home devices, as well as virtual reality synthesis.

Fig. 2.1 depicts the typical scenario of acoustic speaker localization and tracking. In an enclosed environment (e.g. reverberant room), several speakers move and talk, the microphone array recordings are used to extract the locations and sounds of respective speakers. Three apparent challenges arise in this problem, i.e. 1) due to the sound reflections at surfaces of obstacles, the location cues of speakers are ambiguous; 2) due to the movement of speakers, there are only limited sound recordings available corresponding to a certain speaker at a certain location; and 3) due to the time-varying speaker states, there are difficulties in filtering and associating location and speech estimates with respective speakers, and hence forming correctly connected location trajectories and sound streams over time.

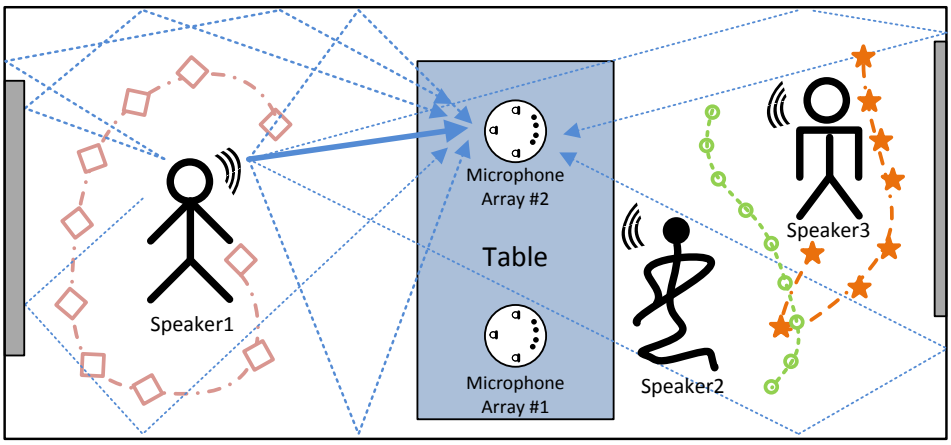

Figure 2.1: Pictorial representation for the studied problem of localization and tracking multiple moving speakers using microphone arrays in a reverberant enclosure. For the clarity of the figure, only some of possible sound paths from Speaker 2 to Microphone Array #2 are drawn. The solid line shows direct-path, while the dotted lines represent sound reflection paths. Track of speakers are denoted with $\diamondsuit$, $\circ$ and $\star$, respectively.

This thesis addresses the challenges in a systematic approach. As shown in Fig. 2.2, speech signals are modeled as the input to the system composed of the room, microphones and the speaker locations. The system produces microphone recordings for signal processing.

There are in general two steps in solving the problem, depending on the treatment of the speaker locations[1]. The first step estimates speaker locations in snapshots (cf. Section 2.5). The second step focuses on the dynamical feature of the system as speaker locations vary over time (cf. Section 2.6).

Following the signal and system representation in Fig. 2.2, models for speech signals, room acoustics, microphone recordings, and theories for the localization and tracking are studied and discussed in the listed order.

[1]This is actually a rather archetypal approach in the estimation paradigm.

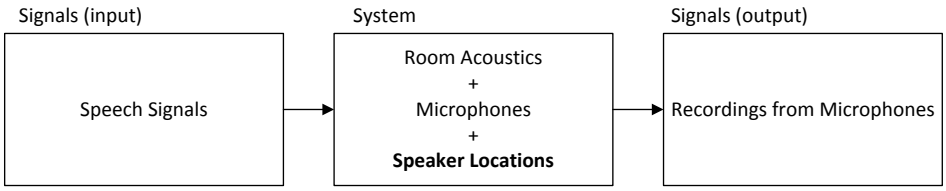

Figure 2.2: Signal and system representation of the studied scenario.

2.2 Speech Signals

Human speech is a dynamic and informative type of signal widely used in daily communications. Corresponding to the audible frequency range of about 20Hz to 20kHz for human hearing, speech signals are commonly regarded as both wideband and bandlimited. Moreover, beside parametric models, speech signals are often treated as realizations of stochastic processes, and wide sense stationary (WSS) and ergodic assumptions are usually made (with cautions) to assist the analysis and engineering processing.

2.2.1 Stochastic Assumptions

Strictly speaking, the process of speech production is nonstationary, i.e., not strong sense stationary (SSS) [47–49]. The often applied WSS and jointly-WSS assumptions[2] imply that the mean and correlation functions are translation-invariant [47]. Moreover, since it is the realizations (speech signals) rather than the ensemble, that are available in practice, the speech production process is assumed correlation ergodic (and hence mean-ergodic). Thus the ensemble averages are estimated by the temporal averages (mean and correlations) [47, 48], i.e.

$$\mu_s \triangleq \mathbb{E}[s(t)] = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} s(t)dt, \tag{2.1}$$

[2]Since speech processing is an applied discipline, the trade-off between theoretical rigorousness and practical applications is often unavoidable. See e.g. [48] for the philosophical discussions on this matter.

$$R_s(\tau) \triangleq \mathbb{E}[s(t_1) \cdot s(t_2)] = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} s(t+\tau) \cdot s(t) dt, \tag{2.2}$$

$$R_{s_i s_j}(\tau) \triangleq \mathbb{E}[s_i(t_1) \cdot s_j(t_2)] = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} s_i(t+\tau) \cdot s_j(t) dt, \tag{2.3}$$

where $\mathbb{E}[\cdot]$ denotes mathematical expectation, and $s(t), s_i(t_1), s_j(t_2) \in \mathbb{R}$ denote the random variables of processes (e.g. speech production) at time $t, t_1, t_2 \in \mathbb{R}$ respectively. μ_s, $R_s(\tau)$ and $R_{s_i s_j}(\tau)$ denote the mean, autocorrelation and cross-correlation respectively, $i \neq j$ are indices, and time difference $\tau \triangleq t_1 - t_2$. Moreover, for zero-mean processes, the correlations equal to the corresponding covariances. For microphone array processing, the covariance matrix is frequently used, e.g. in subspace methods.

The power spectral density (PSD) and cross-power spectral density (CSD) are respectively defined as [47, 49]

$$G_s(\Omega) \triangleq \lim_{T\to\infty} \int_{-\frac{T}{2}}^{\frac{T}{2}} R_s(\tau) e^{-\jmath\Omega\tau} d\tau, \tag{2.4}$$

$$G_{s_i s_j}(\Omega) \triangleq \lim_{T\to\infty} \int_{-\frac{T}{2}}^{\frac{T}{2}} R_{s_i s_j}(\tau) e^{-\jmath\Omega\tau} d\tau, \tag{2.5}$$

where $\jmath \triangleq \sqrt{-1}$, and Ω is the angular frequency in unit of radian per second. Indeed, (2.4) and (2.5) take the form of the Fourier transform (assuming convergence), and are often called the Wiener-Khinchin relations.

Speech typically remains stationary for ranges of only tens of milliseconds [48]. Consequently, the long term properties (e.g. the correlations and spectral densities) are estimated using short-term observations, and the integral intervals in (2.1) through (2.5) are replaced with short time periods. Moreover, for concurrent speakers, the time-frequency (TF) sparsity [50] is often applied, assuming that signals from different speakers do not significantly overlap spectrotemporally. Since it is usually the microphone signals that are available in practice, further discussions on short-time processing of signals will appear in Section 2.4.

2.2.2 Parametric Models

The speech signal model can also be parametrized based on certain knowledge of the speech production mechanism. Simply speaking, human speech is produced by physically structured airflow from the respiratory system. Based on the source excitation - vocal tract models for the mechanism of speech production [48, 51], resulting speech sounds are mainly composed of the voiced and unvoiced sounds, i.e.

$$s(t) = \begin{cases} \mathfrak{g}(t) * \mathfrak{v}(t), \text{ if voiced} \\ \mathfrak{n}(t) * \mathfrak{v}'(t), \text{ if unvoiced,} \end{cases} \tag{2.6}$$

where the operator $*$ denotes the convolution, $\mathfrak{g}(t)$ the glottal waveform [52], $\mathfrak{n}(t)$ the random excitation noise, and $\mathfrak{v}(t)$ and $\mathfrak{v}'(t)$ the vocal tract impulse responses, which can be modeled as time-varying autoregressive (AR) or autoregressive-moving-average (ARMA) systems [53]. The unvoiced sounds are typically stochastic waveforms, while the voiced sounds are harmonically structured and generally dominant in the signal power.

Moreover, inferences of psychoacoustic studies are also often applied in modeling the human auditory system [46, 54]. Typical examples include the critical bands [55, 56] that characterize the fact that the ability of the human cochlea in distinguishing between individual frequency tones varies as a function of frequency, as well as the auditory filters (e.g. the gammatone and gammachirp filters etc.) that mimic the frequency response [57, 58].

2.3 Room Acoustics

In the typical scenario, speech sounds produced by the speaker(s) cause air pressure changes, which are propagated through air and then transduced into electrical signals by the microphone(s) for subsequent processing (e.g. signal conditioning, analog-to-digital conversion, and digital signal processing algorithms, etc.).

In an enclosed room with dimensions relatively large compared to the

sound wavelength, the sound waves reflected at wall surfaces can induce diffuse sound fields. The acoustic response of the enclosed room between a sound source and a receiver, is usually expressed as the superposition of direct-path signals and all its reflections, which can be modeled as a linear time-invariant (LTI) causal system with memory, for the source and receiver with fixed locations[3]. The room transfer function (RTF) and room impulse response (RIR) models are discussed as follows.

2.3.1 Room Transfer Function

Assuming homogeneous, isotropic and lossless medium, following the wave theory (cf. Appendix A), the sound wave propagation at the steady state can be described with a homogeneous partial differential equation (PDE)

$$\frac{1}{v^2}\frac{\partial^2 \mathrm{p}(\vec{\wp}_i, t)}{\partial t^2} = \nabla^2 \mathrm{p}(\vec{\wp}_i, t), \tag{2.7}$$

where v is the velocity of sound propagation, $\mathrm{p}(\vec{\wp}_i, t)$ the sound pressure at location $\vec{\wp}_i$, and operator ∇^2 the Laplacian.

In presence of a single frequency point source, the sound field induced can then be found by modifying (2.7) into the inhomogeneous PDE

$$\frac{1}{v^2}\frac{\partial^2 \mathrm{p}(\vec{\wp}_i, t)}{\partial t^2} - \nabla^2 \mathrm{p}(\vec{\wp}_i, t) = s(\vec{\wp}_q, t), \tag{2.8}$$

where $s(\vec{\wp}_q, t)$ denotes the point sound source at location $\vec{\wp}_q$.

The PDE (2.8) can be converted to the Helmholtz equation and solved via the Green's function, which for free space gives the transfer function [59, 60]

$$\mathrm{H}_{qi}^{(\mathrm{fs})}(\Omega) \triangleq \frac{\mathrm{P}(\vec{\wp}_i, \Omega)}{S(\vec{\wp}_q, \Omega)} = \frac{e^{-\jmath k_\lambda r_{qi}}}{4\pi r_{qi}}, \tag{2.9}$$

where $k_\lambda \triangleq \Omega/v = 2\pi/\lambda$, and λ is the wavelength of the sound signal. $\mathrm{P}(\vec{\wp}_i, \Omega)$ and $S(\vec{\wp}_q, \Omega)$ are Fourier transforms of $\mathrm{p}(\vec{\wp}_i, t)$ and $s(\vec{\wp}_q, t)$ respec-

[3] Ignoring the effects such as air pressure change due to temperature variations over time.

tively, $\vec{\wp}_q \triangleq [x_q, y_q, z_q]$, $\vec{\wp}_i \triangleq [x_i, y_i, z_i]$, and $r_{qi} \triangleq \|\vec{\wp}_i - \vec{\wp}_q\|$ the distance between the source and location $\vec{\wp}_i$. In the far field where r_{qi} is sufficiently large, the spherical wavefront described in (2.9) is often treated as planar.

In an empty enclosed rectangular room with rigid surfaces (room dimensions L_x, L_y and L_z and volume V), a sound source can generate resonances (i.e. room modes). The RTF can then be represented by [59, 61]

$$\mathrm{H}_{qi}^{(\mathrm{rm})}(\Omega) = \sum_j \frac{\mathrm{P}_j(\vec{\wp}_i)\mathrm{P}_j(\vec{\wp}_q)}{V \cdot (k_\lambda^2 - \|\vec{k}_j\|^2)} \tag{2.10}$$

where $\vec{k}_j = \pi[\frac{n_x}{L_x}, \frac{n_y}{L_y}, \frac{n_z}{L_z}]$ is the eigenvalue corresponding to the j-th eigenfrequency, where integers n_x, n_y and n_z are the number of nodal planes perpendicular to respectively the x-axis, y-axis and z-axis of the room [59, 61], and $\mathrm{P}_j(\vec{\wp}_i)$ is the orthogonal eigenfunction dependent on the room boundaries,

$$\mathrm{P}_j(\vec{\wp}) = \frac{1}{8}\sum_{\ell=1}^{8} e^{-\jmath \vec{k}_j \cdot \vec{\wp}_\ell}, \tag{2.11}$$

where $\vec{\wp} \triangleq [x, y, z]$, $\{\vec{\wp}_\ell\}_{\ell=1}^8 \triangleq \{[\pm x, \pm y, \pm z]\}$, and $\vec{k}_j \cdot \vec{\wp}_\ell$ is a dot product.

The image source method (ISM) [59, 62] is often used to numerically simulate room acoustics in the geometrical manner. It uses the RTF model (2.9) and represents the sound reflection of a point source at a reflecting wall as the direct-path from an image source at a location mirrored by that wall. Then each image source is again imaged by other walls, and the sound pressure at a certain location is the superposition of the contributions from all sources. The rigid-boundary RTF can then be expressed as [59]

$$\mathrm{H}_{qi}^{(\mathrm{ism})}(\Omega) = \sum_{r=-\infty}^{\infty}\sum_{p=1}^{8} \frac{e^{-\jmath(\Omega/v)\|R_q - R_r\|}}{4\pi\|R_q - R_r\|}, \tag{2.12}$$

where $\{R_q\}_{p=1}^8 \triangleq \{[x_q \pm x_i, y_q \pm y_i, z_q \pm z_i]\}$ is given by permutations over $\pm$, and $\{R_r\} \triangleq \{2[n_1 L_x, n_2 L_y, n_3 L_z],\ \forall\ n_1, n_2, n_3 \in \mathbb{Z}\}$.

Schroeder also carried out some pioneering studies on the statistical prop-

erties of the diffuse field [63–65]. The Schroeder frequency specifies the lowest sound frequency to have high modal overlap, i.e.

$$f_{schroeder} \approx 2000\sqrt{\frac{T_{60}}{V}}, \tag{2.13}$$

where T_{60} denotes the reverberation time[4]. The mixing time specifies the time required for at least 10 reflections overlap within a characteristic time resolution of 24ms, i.e. $t_{mix} \approx \sqrt{V}$. Diffuse field prevails the room after t_{mix} following an impulse, for signals of frequencies above $f_{schroeder}$. Thus for example, a room of $V = 100\text{m}^3$ and $T_{60} = 1\text{s}$ corresponds to the Schroeder frequency of $f_{schroeder} \approx 200\text{Hz}$, and the mixing time of $t_{mix} \approx 10\text{ms}$. Other general statistical characteristics e.g. echo density, modal density, and peak density etc. can be found in [63–65].

2.3.2 Room Impulse Response

The time domain RIR can be converted from the frequency domain RTF via the inverse Fourier transform (IFT) (or vice versa), i.e.

$$\mathrm{h}_{qi}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{H}_{qi}(\Omega)e^{\jmath\Omega t}d\Omega. \tag{2.14}$$

For example, in free space, from (2.9) the impulse response becomes

$$\mathrm{h}_{qi}^{(\mathrm{fs})}(t) = \frac{\delta(t-\frac{r_{qi}}{v})}{4\pi r_{qi}}, \tag{2.15}$$

where $\delta(\cdot)$ is the Dirac delta function.

Using the ISM, from (2.12) the rigid-boundary RIR becomes [59]

$$\mathrm{h}_{qi}^{(\mathrm{ism})}(t) = \sum_{r=-\infty}^{\infty}\sum_{p=1}^{8} \frac{\delta(t-\frac{\|R_q-R_r\|}{v})}{4\pi\|R_q-R_r\|}. \tag{2.16}$$

[4]Reverberation of an enclosed environment is usually characterized with T_{60}, which is the time required for sound to decay by 60dB.

It is obvious that the RIR is a superposition of impulse responses from the direct-path and reflections, depending on the source and sensor locations. It hence represents an LTI system for fixed source and sensor locations.

For simulating non-rigid room boundaries, reflection constants due to surfaces can be further included in the models (2.12) and (2.16) [59, 62]. By neglecting the air attenuation and assuming that the reflection coefficient (denoted as $\beta_r \in \mathbb{C}$) of all surfaces are the same, the Eyring's formula [61,66] describes the relationship between the reverberation time and β_r, i.e.

$$\beta_r = e^{-\frac{6\ln 10}{v(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z})T_{60}}}. \tag{2.17}$$

In practice, the exact value of the reverberation time T_{60} of the environment is often unknown *a priori*, but it can be measured [67] or estimated from sound recordings [68]. Intuitively, the RIR can be categorized into three parts, viz. the direct-path, early reflections and the late (diffuse) reverberation [69, 70]. Fig. 2.3 shows an example of the RIR.

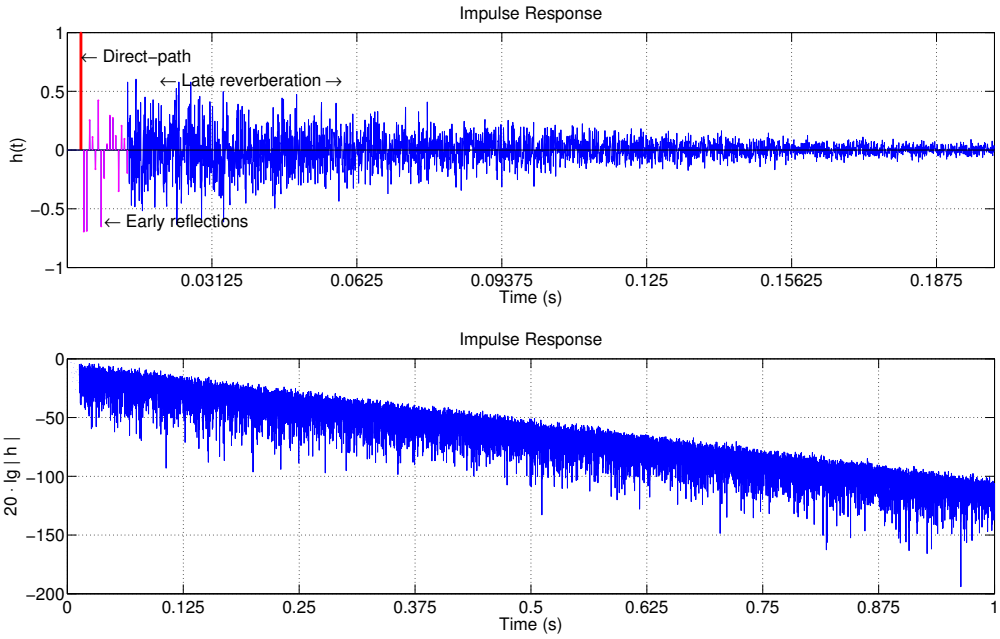

Figure 2.3: An example of the room impulse response, plotted in linear scale (top panel) and logarithmic scale (bottom panel).

2.4 Microphone Array Signals

As shown in Fig. 2.1 and Fig. 2.2, microphones are placed in the room and constitute part of the system producing sound recordings, which are availed for subsequent processing. Recordings from the microphone array(s) bear the desired information, e.g. sound signals (waveforms) from sources, and their locations. Thus this thesis assumes that the dimensions of the microphone array are not too large, microphones are lossless omnidirectional, and are located on the same plane of speakers. Moreover, the far-field assumption is made that the distance between sources and microphones are sufficiently large. Thus from a particular source to the microphones, the gain constant as in (2.9) and (2.15) can be normalized, and the direction is regarded the same.

2.4.1 Array Signal Denotations

The signals impinging the microphone array are denoted in time domain as

$$\mathbf{x}(t) \triangleq [x_1(t), \cdots, x_i(t), \cdots, x_{I_M}(t)]^T, \tag{2.18}$$

where integer I_M denotes the total number of microphones, and $x_i(t)$ denotes the time domain signal at the microphone with index i, $i = 1, \cdots, I_M$.

The frequency domain denotation for the microphone array signal is thus the Fourier transform of (2.18)

$$\mathbf{X}(\Omega) \triangleq [X_1(\Omega), \cdots, X_i(\Omega), \cdots, X_{I_M}(\Omega)]^T, \tag{2.19}$$

where $X_i(\Omega)$ denotes the frequency domain signal at microphone i.

Anechoic and reverberant models are developed respectively as follows.

2.4.2 Anechoic Case

TIME DOMAIN

Consider first the anechoic case where some speakers located in the far-field of the microphones, and the maximum aperture of the microphone array is small compared to its bulk distance to the source. The spherical wavefronts (2.8) can thus be regarded as planar in the far-field. Assuming that the transmission medium is at rest, homogeneous and lossless (i.e. no reflection, diffusion, refraction or absorption), thus the superposition of plane waves impinging microphones and the additive noise can be expressed as

$$x_i(t) = \sum_{q=1}^{Q} s_q(t - t_{d_{qi}}) + n_i(t), \tag{2.20}$$

where $n_i(t) \in \mathbb{R}$ is the additive noise, Q is the total number of concurrent speakers, and $t_{d_{qi}}$ is the direct-path time delay for the sound to travel from speaker q to microphone i, which is a function of the direction of arrival (DOA) θ_q, i.e.

$$t_{d_{qi}}(\theta_q) = \frac{\overrightarrow{m}_i \cdot \overrightarrow{\wp}_q(\theta_q)}{v\|\overrightarrow{\wp}_q(\theta_q)\|}, \quad \overrightarrow{\wp}_q(\theta_q) \triangleq \|\overrightarrow{\wp}_q\| e^{\jmath\theta_q}, \tag{2.21}$$

where the denominator is a dot product, $\overrightarrow{m}_i$ and $\overrightarrow{\wp}_q$ denotes the locations of the microphone i and the speaker q with respect to the origin of the coordinate system (e.g. the center of gravity of the microphone array).

The signals received by the array are represented by stacking up those by each microphone

$$\mathbf{x}(t) = [\sum_{q=1}^{Q} s_q(t - t_{d_{q1}}), \cdots, \sum_{q=1}^{Q} s_q(t - t_{d_{qI_M}})]^T + \mathbf{n}(t), \tag{2.22}$$

where

$$\mathbf{n}(t) \triangleq [n_1(t), \cdots, n_{I_M}(t)]^T. \tag{2.23}$$

FREQUENCY DOMAIN

The simple multi-source model in (2.20) via the Fourier transform becomes

$$X_i(\Omega) = \sum_{q=1}^{Q} e^{-\jmath\Omega t_{d_{qi}}} \cdot S_q(\Omega) + N_i(\Omega), \tag{2.24}$$

where $S_q(\Omega)$ and $N_i(\Omega) \in \mathbb{C}$ are respectively the Fourier transforms of $s_q(t)$ and $n_i(t)$.

In this case, signals received by the array in (2.22) becomes,

$$\mathbf{X}(\Omega) = \mathbf{D}(\boldsymbol{\theta}, \Omega)\mathbf{S}(\Omega) + \mathbf{N}(\Omega), \tag{2.25a}$$

where,

$$\mathbf{S}(\Omega) = [S_1(\Omega), \cdots, S_q(\Omega), \cdots, S_Q(\Omega)]^T, \tag{2.25b}$$
$$\mathbf{N}(\Omega) = [N_1(\Omega), \cdots, N_i(\Omega), \cdots, N_{I_M}(\Omega)]^T, \tag{2.25c}$$

and

$$\mathbf{D}(\boldsymbol{\theta}, \Omega) \triangleq [\mathbf{d}(\theta_1, \Omega), \cdots, \mathbf{d}(\theta_q, \Omega), \cdots, \mathbf{d}(\theta_Q, \Omega)], \tag{2.26}$$

$$\mathbf{d}(\theta_q, \Omega) \triangleq \left[\mathrm{e}^{-\jmath\Omega t_{d_{q1}}(\theta_q)}, \ldots, \mathrm{e}^{-\jmath\Omega t_{d_{qi}}(\theta_q)}, \ldots, \mathrm{e}^{-\jmath\Omega t_{d_{qI_M}}(\theta_q)}\right]^T, \tag{2.27}$$

the vector parameter is defined as

$$\boldsymbol{\theta} \triangleq [\theta_1, \cdots, \theta_q, \cdots, \theta_Q], \tag{2.28}$$

which is unknown and to be estimated from observed signals $\mathbf{X}(\Omega)$.

2.4.3 Reverberant Case

TIME DOMAIN

In more challenging conditions, e.g. room reverberation and moving sources, the simple model in (2.20) may not suffice. Thus using the RIR (2.14), a

more complete model is

$$x_i(t) = \sum_{q=1}^{Q} \mathrm{h}_{qi}(t) * s_q(t) + n_i(t), \tag{2.29}$$

and the array signal is denoted as

$$\mathbf{x}(t) = \mathbf{h}(t) * \mathbf{s}(t) + \mathbf{n}(t), \tag{2.30}$$

where

$$\mathbf{h}(t) \triangleq \begin{bmatrix} \mathrm{h}_{11}(t) & \cdots & \mathrm{h}_{Q1}(t) \\ \vdots & \ddots & \vdots \\ \mathrm{h}_{1I_M}(t) & \cdots & \mathrm{h}_{QI_M}(t) \end{bmatrix}, \tag{2.31}$$

$$\mathbf{s}(t) \triangleq [s_1(t), \cdots, s_Q(t)]^T, \tag{2.32}$$

the term $\mathbf{h}(t) * \mathbf{s}(t)$ of (2.30) follows matrix multiplication, except that the element-wise operator $*$ denotes convolution. It depends on the locations of the speaker and microphone array (e.g. θ_q) for a given environment, and is hence time-varying for a moving speaker.[5] When acoustic reflections and absorptions are negligible, (2.30) simplifies to (2.22).

FREQUENCY DOMAIN

Corresponding to (2.29), the frequency domain signal can be expressed as

$$X_i(\Omega) = \sum_{q=1}^{Q} \mathrm{H}_{qi}(\theta_q, \Omega) \cdot S_q(\Omega) + N_i(\Omega). \tag{2.33}$$

Thus corresponding to (2.30), the array signal in frequency domain is

$$\mathbf{X}(\Omega) = \mathbf{H}(\boldsymbol{\theta}, \Omega)\mathbf{S}(\Omega) + \mathbf{N}(\Omega), \tag{2.34}$$

[5]The same notation $\mathbf{x}(t)$ is used in (2.18) and (2.30) for notational simplicity, because for a specific algorithm, either model (2.18) or (2.30) is used (not both). This is also the case in (2.25) and (2.34). Unless otherwise specified, (2.30) and (2.34) are used in what follows.

where

$$\mathbf{H}(\boldsymbol{\theta},\Omega) \triangleq \begin{bmatrix} \mathrm{H}_{11}(\theta_1,\Omega) & \cdots & \mathrm{H}_{Q1}(\theta_Q,\Omega) \\ \vdots & \ddots & \vdots \\ \mathrm{H}_{1I_M}(\theta_1,\Omega) & \cdots & \mathrm{H}_{QI_M}(\theta_Q,\Omega) \end{bmatrix}, \tag{2.35}$$

Entry $\mathrm{H}_{qi}(\theta_q,\Omega)$ in $\mathbf{H}(\boldsymbol{\theta},\Omega)$ is the frequency domain transfer function from source q to sensor i (also the Fourier transform of $\mathrm{h}_{qi}(t)$), which is dependent on the locations of the source and sensor array. When acoustic reflections and absorptions are negligible, (2.35) simplifies to (2.25).

2.4.4 Discrete and Short-time Processing

Discrete Time Samples

Note that in practice, the continuous time domain signal $x_i(t)$ is sampled (discretized) at a certain frequency f_s before processing, and the sampled signals at time $t = n/f_s = n \cdot T_s$ ($n \in \mathbb{Z}$) is denoted as $x_i[n]$, i.e.

$$x_i[n] \triangleq x_i(n \cdot T_s), \tag{2.36}$$

where $T_s = 1/f_s$ is the sampling period.

Discrete Short-time Fourier Transform

Moreover, in practical implementations, discrete time speech signals (e.g. (2.36)) are often processed in short time intervals using analysis - synthesis techniques. The discrete Fourier transform (DFT), or more prevalently, the short-time Fourier transform (STFT) (see e.g. [71–73]) is used instead of the Fourier transform.

Here for a snapshot (denoted frame k) of sampled time domain signal, the STFT coefficient for normalized frequency $\omega = \frac{\Omega}{f_s}$ is

$$X_i(k,\omega) = \sum_{n=kM-N+1}^{kM} x_i[n]w[kM-n]e^{-\jmath\omega n}, \tag{2.37}$$

where k is the time frame index for the N samples, integer M is shifting step

size, and real-valued $w[\cdot]$ is the (sliding) analysis window function. Due to the finite time length, ω is sampled with $\omega_m = \frac{2\pi}{N}m,\ m = 0, 1, \cdots, N-1$.

The DFT can be regarded as a special case of the STFT when the rectangular window is used. Window functions were originally used to suppress spectral "leakage" caused by the rectangular window. Typical window functions include the Hann, Hamming, Blackman windows and many more, e.g.

$$w[n] = \begin{cases} a - (1-a)\cos\left(\frac{2\pi n}{N-1}\right),\ n \in [0, N-1] \\ 0,\ \text{otherwise} \end{cases}, \tag{2.38}$$

where the Hann window has $a = 0.5$, and the Hamming window has $a \approx 0.54$. Typical window length is about 30ms for many speech signal processing algorithms, assuming stationarity in short time intervals. A summary of window functions can be found in [74]. In what follows and unless otherwise noted, STFT notations (2.37) are used for in frequency domain processing, and the time frame index may be suppressed for notational simplicity.

2.5 Localization Overview

As shown in the general signal models (2.30) and (2.34), the speaker locations (e.g. $\boldsymbol{\theta}$) are embedded in, and hence can be found from observed signals $\mathbf{x}(t)$ (or the sampled signals $\mathbf{x}[n]$) and $\mathbf{X}(\Omega)$ (or the STFT $\mathbf{X}(k, \omega)$), respectively.

This section provides an overview of some existing state-of-the-art speaker localization methods that are representative and have been implemented in this thesis. Because of the significant body of literature of the localization methods, it is impractical to go through exhaustively all existing works. However, it is necessary to review the motivations and assumptions of existing state-of-the-art techniques, and implement them for comparative studies. In what follows, the localization methods are classified into three groups, viz. steered response power beamformers, subspace methods and TDOA-based methods. Note that each category has a vast range of members and variants.

2.5.1 Wideband Beamformers

The most intuitive localization method is to scan the location space using beamformers. Wideband beamformers are required for speaker localization.

WIDEBAND BEAMFORMERS

A good variety of wideband beamformers can be found in the literature (see e.g. [31, 75, 76] and the references therein). A simplified fixed wideband WLS beamformer that uses a finite impulse response (FIR) "filter-and-sum" structure (based on [31, Chapter 4.2]) is implemented in the thesis for speech separation. Definitions and the formulation are summarized as follows.

$\check{\mathbf{d}}(\theta,\omega)$ is the wideband beamformer steering vector defined as

$$\check{\mathbf{d}}(\theta,\omega) \triangleq \left[1,\ldots,e^{-j\omega\cdot jT_s},\ldots,e^{-j\omega\cdot(J-1)T_s}\right]^T \otimes \mathbf{d}(\theta,\omega) \quad , \tag{2.39}$$

where $\otimes$ is the Kronecker tensor product, θ the DOA, and the array steering vector $\mathbf{d}(\theta,\omega)$ is defined in (2.27).

$P(\theta,\omega)$ is the beamformer response defined as

$$P(\theta,\omega) \triangleq \mathbf{w}^H \cdot \check{\mathbf{d}}(\theta,\omega) \tag{2.40}$$

where $\mathbf{w} \in \mathbb{R}^{J\cdot I_M\times 1}$ is the weight vector to be solved (depending on the desired DOA of speakers as well as the passband and stopband frequencies), and J the number of taps of the FIR filter.

$P_d(\theta,\omega)$ is the desired beamformer response, e.g. for simplicity,

$$P_d(\theta,\omega) \triangleq \begin{cases} e^{-j\omega\tau_0}, \text{ for } \theta \in \Theta_{ml} \wedge \omega \in \Omega_{pb}/f_s \\ 0, \text{ for } \theta \in \Theta_{sl} \vee \omega \in \Omega_{sb}/f_s \end{cases} \tag{2.41}$$

where Ω_{pb} and Ω_{sb} are the passband and the stopband frequency ranges respectively. $\Omega_{pb} = 2\pi[20,4000]$Hz, $\Omega_{sb} = 2\pi(4000,8000]$Hz. $\tau_0 = f_s \cdot 10$s (chosen for the far-field assumption). Θ_{ml} is the angle range of the

beamformer mainlobe, and Θ_{sl} the angle range of the beamformer sidelobe.

For the WLS wideband beamformer, the desired weight vector $\mathbf{w}$ is derived by formulating a weighted least square problem [31]

$$\min_{\mathbf{w}} \mathcal{J}_{LS}(\mathbf{w}) \triangleq \min_{\mathbf{w}} \int_{\Omega_R} \int_{\Theta} \nu(\theta,\omega)|P(\theta,\omega) - P_d(\theta,\omega)|^2 d\theta d\omega \tag{2.42}$$

where $\Omega_R = 2\pi[20, 8000]\text{Hz}/f_s$ is the frequency range of interest, and $\Theta \in [0, 360^\circ)$ the range of DOAs. $\nu(\theta,\omega)$ is the weighting function defined as

$$\nu(\theta,\omega) \triangleq \begin{cases} \alpha, \text{ for } \theta \in \Theta_{ml} \wedge \omega \in \Omega_{pb} \\ 1-\alpha, \text{ for } \theta \in \Theta_{sl} \vee \omega \in \Omega_{sb} \end{cases} \tag{2.43}$$

where $\alpha \in (0,1)$ is a weighting parameter.

Based on the knowledge that $\mathcal{J}_{LS}(\mathbf{w}) \in \mathbb{R}$, (2.42) simplifies to (see [31])

$$\mathcal{J}_{LS}(\mathbf{w}) = \mathbf{w}^H \mathbf{G}_{ls} \mathbf{w} - 2\mathbf{w}^H \mathbf{g}_{ls} + g_{ls}, \tag{2.44}$$

where $[\cdot]^H$ denotes the Hermitian (conjugate transpose) operation,

$$\mathbf{G}_{ls} = \alpha \int_{\Omega_{pb}} \int_{\Theta_{ml}} \check{\mathbf{D}}_R(\theta,\omega) d\theta d\omega + (1-\alpha) \int_{\Omega_{pb}} \int_{\Theta_{sl}} \check{\mathbf{D}}_R(\theta,\omega) d\theta d\omega \tag{2.45}$$

$$\mathbf{g}_{ls} = \alpha \int_{\Omega_{pb}} \int_{\Theta_{ml}} \Big(\check{\mathbf{d}}_R(\theta,\omega)\cos(\tau_0\omega) - \check{\mathbf{d}}_I(\theta,\omega)\sin(\tau_0\omega) \Big) d\theta d\omega \tag{2.46}$$

$$\check{\mathbf{D}}(\theta,\omega) = \check{\mathbf{d}}(\theta,\omega)\check{\mathbf{d}}(\theta,\omega)^H, \tag{2.47}$$

$\check{\mathbf{D}}_R = \Re\{\check{\mathbf{D}}\}$, $\check{\mathbf{D}}_I = \Im\{\check{\mathbf{D}}\}$. $\check{\mathbf{d}}_R = \Re\{\check{\mathbf{d}}\}$, $\check{\mathbf{d}}_I = \Im\{\check{\mathbf{d}}\}$. $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real and image part respectively.

Since $\frac{\partial \mathcal{J}_{LS}(\mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{G}_{ls}\mathbf{w} - 2\mathbf{g}_{ls}$, the optimal weight vector solution of (2.42) is

$$\mathbf{w}_{opt} = \mathbf{G}_{ls}^{-1} \mathbf{g}_{ls}. \tag{2.48}$$

Thus the beamformer output is

$$y[n] = \mathbf{w}_{opt}^H \mathbf{x}_{BF}[n], \tag{2.49}$$

where

$$\mathbf{x}_{BF}[n] = \left[\mathbf{x}_1[n],\ldots,\mathbf{x}_i[n],\ldots,\mathbf{x}_{I_M}[n]\right]^T, \tag{2.50}$$

$$\mathbf{x}_i[n] = \left[x_i[n],\ldots,x_i[n+j],\ldots,x_i[n+J-1]\right]. \tag{2.51}$$

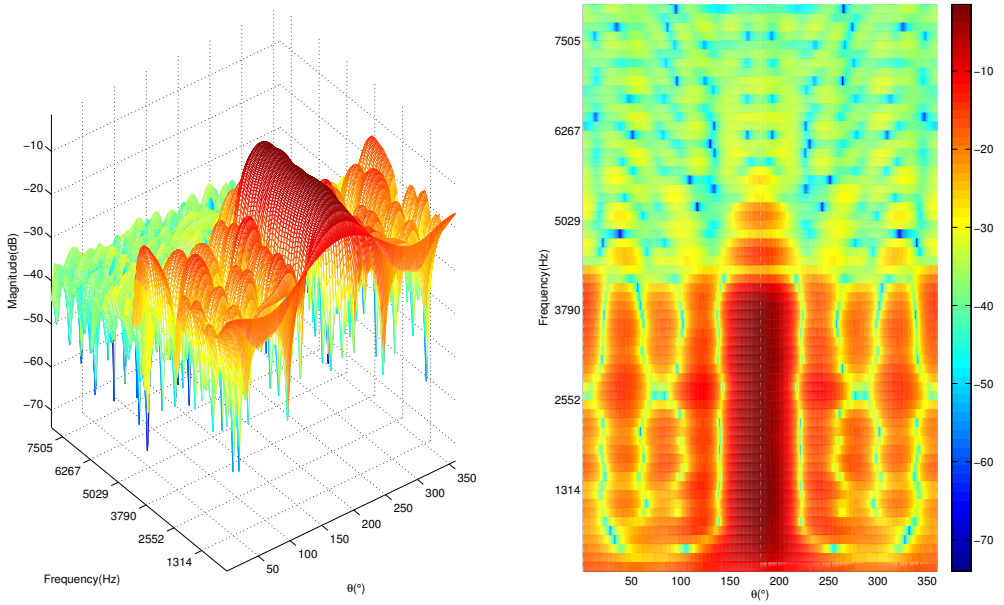

Figure 2.4: An example of the WLS beamformer response.

The spatial and spectral response (i.e. $\hat{P}_d(\theta,\omega) = \mathbf{w}_{opt}^H \check{\mathbf{d}}(\theta,\omega)$) of the designed WLS wideband beamformer is shown in Fig. 2.4. The cut-off frequency is 4000Hz, the mainlobe points at 180°, and the sidelobe covers the range outside of $180° \pm 15°$. A UCA with 8 sensors and a diameter of 0.1m is used.

Similar to the WLS beamformer, a variety of fixed beamformers can be designed by formulating different topologies, e.g. delay-and-sum, and various cost functions and constraints [77–80]. Besides, adaptive beamformers such as linearly constrained minimum variance (LCMV) [81] and generalized sidelobe canceller (GSC) [82], etc. can also be found in the literature ([31,83] provide good summaries of wideband beamformer design). On the one hand, the wideband beamformer steers the beam over possible DOAs and

can be used to find the maximal response power for the purpose of source localization. On the other hand, the beamformers can also be used to extract source signals from given DOAs. Therefore the wideband beamformer is an estimator for both DOA and waveform parameters.

EIGEN-BEAMFORMER

Eigenbeam techniques have been recently developed for some special sensor array apertures, e.g. the spherical arrays [84, 85] and the uniform circular arrays (UCA) [32–34]. The main motivation is to use the spherical or circular symmetry of the sensor array and process the far-field signal in terms of the phase mode excitations. The time-frequency circular harmonics beamformer (TF-CHB) based on [34] is implemented and summarized as follows.

Assuming a plane wave impinging an unbaffled continuous circular aperture with a radius of r_a from DOA θ_{in} (cf. (2.21)), the sound pressure at θ can be expressed as

$$P(k_\lambda r_a, \theta) = P_0 \cdot e^{\jmath k_\lambda r_a \cos(\theta - \theta_{in})}, \tag{2.52}$$

which can be expanded into the phase modes (or circular harmonics), i.e. [34, 86, 87]

$$P(k_\lambda r_a, \theta) = \sum_{\mathsf{p}=-\infty}^{\infty} C_\mathsf{p}(k_\lambda r_a, \theta_{in}) e^{\jmath \mathsf{p}\theta} \overset{\text{truncate}}{\approx} \sum_{\mathsf{p}=-\mathsf{P}}^{\mathsf{P}} C_\mathsf{p}(k_\lambda r_a, \theta_{in}) e^{\jmath \mathsf{p}\theta}, \tag{2.53}$$

where k_λ is the angular wavenumber as defined in (2.9), P_0 the impinging wave, the coefficients $C_\mathsf{p}(k_\lambda r_a, \theta_{in}) \triangleq P_0 \jmath^\mathsf{p} J_\mathsf{p}(k_\lambda r_a) e^{-\jmath \mathsf{p}\theta_{in}}$, the highest order of circular harmonics chosen as $\mathsf{P} = \lceil k_\lambda r_a \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function, and $J_\mathsf{p}(k_\lambda r_a)$ the Bessel function of the first kind of order p. Note that (2.53) can also be interpreted as a spatial Fourier transform [88, 89].

For a circular array with I_M discrete microphones, the spatial Fourier coefficients $C_\mathsf{p}(k_\lambda r_a, \theta_{in})$ are approximated with

$$\tilde{C}_\mathsf{p}(\omega, \theta_{in}) = \frac{1}{I_M} \sum_{i=1}^{I_M} X_i(\omega) e^{-\jmath \mathsf{p}\dot{\theta}_i}, \ \mathsf{p} \in [-\mathsf{P}, \mathsf{P}], \tag{2.54}$$

where $\dot{\theta}_i$ denotes the angular location of microphone i, and $I_M \geq 2\mathsf{P}+1$ applying the sampling theorem.

Thus the modal coefficients of the microphone array are obtained

$$\mathbf{C}_{\mathrm{CH}}(\omega) = \mathbf{C}_{EB}\mathbf{X}(\omega), \tag{2.55}$$

where

$$\mathbf{C}_{EB} = [\mathbf{c}_{EB}(-\mathsf{P}), \cdots, \mathbf{c}_{EB}(\mathsf{p}), \cdots, \mathbf{c}_{EB}(\mathsf{P})]^T \tag{2.56a}$$

$$\mathbf{c}_{EB}(\mathsf{p}) = \frac{1}{I_M}[e^{-\jmath\mathsf{p}\dot{\theta}_1}, \cdots, e^{-\jmath\mathsf{p}\dot{\theta}_i}, \cdots, e^{-\jmath\mathsf{p}\dot{\theta}_{I_M}}]^T. \tag{2.56b}$$

A Tikhonov-regularized filter is used to regularize the responses of the individual eigenbeams,

$$\mathbf{B}_{\mathrm{CHB}} = [B_{-\mathsf{P}}(k_\lambda r_a), \cdots, B_{\mathsf{P}}(k_\lambda r_a)]^T, \tag{2.57}$$

where the coefficients are given as

$$B_{\mathsf{p}}(k_\lambda r_a) = \frac{w^*_{\mathsf{p}}(k_\lambda r_a)}{\|w_{\mathsf{p}}(k_\lambda r_a)\|^2 + \beta}, \tag{2.58}$$

$w_{\mathsf{p}}(k_\lambda r_a) = \jmath^{\mathsf{p}} J_{\mathsf{p}}(k_\lambda r_a)$ and $\beta = 6.5 \times 10^{-4}$ is the regularization coefficient.

The response of the TF-CHB at DOA $\theta \in [0, 2\pi)$ for frequency ω is

$$y_{\mathrm{CHB}}(\theta,\, \omega) = \frac{1}{2\mathsf{P}+1}\mathbf{d}_{\mathrm{CHB}}(\theta)\ (\mathbf{B}_{\mathrm{CHB}} \bigodot \mathbf{C}_{\mathrm{CH}}(\omega)), \tag{2.59}$$

where $\bigodot$ denotes the Hadamard product. The modal steering vector is

$$\mathbf{d}_{\mathrm{CHB}}(\theta) = [e^{-\jmath(-\mathsf{P})\theta}, \cdots, e^{-\jmath(\mathsf{p})\theta}, \cdots, e^{-\jmath(\mathsf{P})\theta}]. \tag{2.60}$$

The overall steered response power (SRP) for a wideband signal is

$$\epsilon^{\mathrm{CHB}}(\theta) = \sum_{\omega} |y_{\mathrm{CHB}}(\theta,\, \omega)|. \tag{2.61}$$

The DOA estimates are then denoted as

$$\hat{\Theta} = \{\hat{\theta}\}, \tag{2.62}$$

where $\hat{\theta}$ corresponds to peaks of the SRP $\epsilon^{\mathrm{CHB}}(\theta)$.

2.5.2 Subspace Methods

Subspace-based localization methods usually rely on the spatial covariance matrix of the array signals. For the zero-mean impinging signals, the covariances are equivalent to the correlations as defined in Section 2.2.1.

From (2.2) and (2.3), the correlation matrix is written as [90]

$$\mathbf{R}_{\mathbf{x}}(\tau) = \mathbb{E}[\mathbf{x}(t+\tau)\mathbf{x}^T(t)], \tag{2.63}$$

which converts to the narrowband spectral density matrices via the Fourier transform. Using (2.25) and assuming uncorrelated speech signals and noises, from (2.63) the spectral density matrix [90, 91] is thus (cf. (2.4) and (2.5))

$$\mathbf{R}_{\mathbf{X}}(\omega) = \mathbf{D}(\omega)\mathbf{P}_S(\omega)\mathbf{D}^H(\omega) + \sigma_N^2(\omega)\mathbf{P}_N(\omega), \tag{2.64}$$

where the STFT denotation is used, $\mathbf{P}_S(\omega)$ and $\mathbf{P}_N(\omega)$ denote the spectral density matrix of source signals and noises respectively. $\mathbf{P}_N(\omega)$ is often assumed known to the algorithm. For the simplicity of discussion hereafter, the noise power is further assumed the same level for all microphones [92]. Thus $\mathbf{P}_N(\omega)$ is replaced with $\mathbf{I}_{I_M}$, which is an $I_M \times I_M$ identity matrix.

Under the ergodic and WSS assumptions for speech signals (cf. Section 2.2.1), the (spatial) covariance matrix is finally calculated as an estimate of (2.64) via frequency domain narrowband snapshots [90, 91], e.g.

$$\mathbf{R}(\omega) = \frac{1}{N}\sum_{k=1}^{N}\mathbf{X}(k,\omega)\mathbf{X}^H(k,\omega). \tag{2.65}$$

Then, the covariance matrix can be factorized into signal and noise subspaces. Its eigenvalues and eigenvectors are obtained via eigendecomposi-

tion, i.e.

$$\mathbf{R}(\omega)\ \mathbf{V}(\omega) = \mathbf{V}(\omega)\ \mathbf{\Lambda}(\omega), \tag{2.66}$$

where $\mathbf{\Lambda}(\omega)$ is a diagonal matrix containing the eigenvalues of $\mathbf{R}(\omega)$ in descending order, columns of matrix $\mathbf{V}(\omega)$ are the corresponding eigenvectors of $\mathbf{R}(\omega)$.

Further assume that the number of active sources $Q < I_M$ and the noises are not too strong. The estimated number of signal sources, $\hat{Q}$, or equivalently the rank of the signal subspace, can be obtained using either *a priori* knowledge or the number of eigenvalues that are greater than a selected threshold. The rank of the noise subspace is $\hat{N} = I_M - \hat{Q}$. Thus the matrix of eigenvectors contains the eigenvectors of signal subspace and noise subspace, i.e. [74, 93]

$$\mathbf{V}(\omega) \triangleq [\mathbf{E}_S(\omega)|\mathbf{E}_N(\omega)], \tag{2.67}$$

where $\mathbf{E}_S(\omega)$ is an $I_M \times \hat{Q}$ matrix, while $\mathbf{E}_N(\omega)$ is an $I_M \times \hat{N}$ matrix. Column vectors of the $\mathbf{E}_S(\omega)$ and $\mathbf{E}_N(\omega)$ correspond to the descending order of eigenvalues, span the signal subspace and noise subspace respectively, and are orthonormal.

MUSIC

The MUSIC (Multiple Signal Classification) method was first developed in [27] for the parameter estimation of narrowband signals.

From (2.64), (2.66) and (2.67),

$$\begin{aligned}\mathbf{R}(\omega)\mathbf{E}_N(\omega) =& \mathbf{D}(\omega)\mathbf{P}_S(\omega)\mathbf{D}^H(\omega)\mathbf{E}_N(\omega) + \sigma_N^2(\omega)\mathbf{E}_N(\omega) && \text{(2.68a)}\\ =& \sigma_N^2(\omega)\mathbf{E}_N(\omega), && \text{(2.68b)}\end{aligned}$$

it can be seen that $\mathbf{D}(\omega)\mathbf{P}_S(\omega)\mathbf{D}^H(\omega)\mathbf{E}_N(\omega) = \mathbf{O}_N$, where $\mathbf{O}_N$ is an $I_M \times \hat{N}$ zero matrix. Thus $\mathbf{D}^H(\omega)\mathbf{E}_N(\omega) = \mathbf{O}_N$.

Moreover, since $\mathbf{V}(\omega)\mathbf{V}^H(\omega) = \mathbf{I}_{I_M}$, and

$$\begin{aligned}\mathbf{V}(\omega)\mathbf{V}^H(\omega) =& [\mathbf{E}_S(\omega)|\mathbf{E}_N(\omega)][\mathbf{E}_S(\omega)|\mathbf{E}_N(\omega)]^H \\ =& \mathbf{E}_S(\omega)\mathbf{E}_S^H(\omega) + \mathbf{E}_N(\omega)\mathbf{E}_N^H(\omega),\end{aligned} \tag{2.69}$$

it is apparent that $\mathbf{E}_S(\omega)\mathbf{E}_S^H(\omega) + \mathbf{E}_N(\omega)\mathbf{E}_N^H(\omega) = \mathbf{I}_{I_M}$.

Therefore, the localization function of the MUSIC method is defined as

$$\epsilon^{\text{MUSIC}}(\theta,\omega) = \frac{\mathbf{d}^H(\theta,\omega)\ \mathbf{d}(\theta,\omega)}{\mathbf{d}^H(\theta,\omega)\ \mathbf{E}_N(\omega)\mathbf{E}_N^H(\omega)\ \mathbf{d}(\theta,\omega)} \tag{2.70a}$$

or equivalently,

$$\epsilon^{\text{MUSIC}}(\theta,\omega) = \frac{\mathbf{d}^H(\theta,\omega)\ \mathbf{d}(\theta,\omega)}{\mathbf{d}^H(\theta,\omega)\ (\mathbf{I}_{I_M} - \mathbf{E}_S(\omega)\mathbf{E}_S^H(\omega))\ \mathbf{d}(\theta,\omega)}, \tag{2.70b}$$

where $\mathbf{d}(\theta,\omega)$ is the steering vector, cf. (2.27).

Scanning $\mathbf{d}(\theta,\omega)$ over the array manifold, then $\hat{Q}$ highest peaks of the localization function are selected, and the corresponding DOAs are the estimated source locations. Performance study of the MUSIC can be found in [92, 94].

For wideband applications [95], the localization function can be expressed as

$$\epsilon^{\text{MUSIC}}_{\text{wideband}}(\theta) = \sum_{\omega} \epsilon^{\text{MUSIC}}(\theta,\omega). \tag{2.71}$$

DOA estimates $\hat{\Theta}$ (defined in (2.62)) can then be found from the peaks.

ESPRIT AND EB-ESPRIT

The ESPRIT (Estimation of Signal Parameter via Rotational Invariance Techniques) method was first developed in [28] as a narrowband signal parameter estimation technique and has made significant impact especially in the localization of signal sources. It reduces the computational complexity compared with the MUSIC by exploiting the shift invariance of the sensor array.

From (2.26), denote the shift-invariant subarrays as

$$\mathbf{D}_X(\omega) = \mathbf{J}_X\mathbf{D}(\omega) \tag{2.72a}$$
$$\mathbf{D}_Y(\omega) = \mathbf{J}_Y\mathbf{D}(\omega). \tag{2.72b}$$

where $\mathbf{J}_X$ and $\mathbf{J}_Y$ are $N_s \times I_M$ ($\hat{Q} < N_s < I_M$) matrices that select shift-invariant subarrays, e.g. for linear displacement,

$$\mathbf{J}_X = [\mathbf{I}_{N_s} \mid \mathbf{O}_S] \tag{2.73a}$$
$$\mathbf{J}_Y = [\mathbf{O}_S \mid \mathbf{I}_{N_s}], \tag{2.73b}$$

and here $\mathbf{O}_S$ is a $N_s \times (I_M - N_s)$ zero matrix.

The shift invariance property of the array manifold implies that $\exists\ \mathbf{\Phi}$,

$$\mathbf{D}_Y(\omega) = \mathbf{D}_X(\omega)\mathbf{\Phi}, \tag{2.74}$$

where $\mathbf{\Phi}$ is a $N_s \times N_s$ diagonal matrix that characterizes the subarray shift.

From (2.26), the columns of the array manifold spans the signal subspace, thus $\exists\ \mathbf{T}_S$, a nonsingular $\hat{Q} \times \hat{Q}$ matrix, so that

$$\mathbf{E}_S(\omega) = \mathbf{D}(\omega)\mathbf{T}_S. \tag{2.75}$$

From (2.67), subarray signal subspaces can be obtained, i.e.

$$\mathbf{E}_X(\omega) = \mathbf{J}_X\mathbf{E}_S(\omega) \tag{2.76a}$$
$$\mathbf{E}_Y(\omega) = \mathbf{J}_Y\mathbf{E}_S(\omega), \tag{2.76b}$$

Therefore, $\mathbf{E}_Y(\omega) = \mathbf{E}_X(\omega)\mathbf{T}_S^{-1}\mathbf{\Phi}\mathbf{T}_S \triangleq \mathbf{E}_X(\omega)\mathbf{\Psi}$. Eigenvalues of $\mathbf{\Psi}$ correspond to elements of $\mathbf{\Phi}$. To solve this overdetermined set of equation, the least squares approach gives $\mathbf{\Psi}_{LS} = [\mathbf{E}_X(\omega)^H\mathbf{E}_X(\omega)]^{-1}\mathbf{E}_X(\omega)^H\mathbf{E}_Y(\omega)$, which may be biased.

The more appropriate total-least squares criterion leads to $\mathbf{\Psi}_{TLS} = -\mathbf{E}_{12}\cdot$

$\mathbf{E}_{22}^{-1}$, where $\mathbf{E}_{12}$ and $\mathbf{E}_{22}$ are $N_s \times N_s$ submatrices of matrix $\mathbf{E}$

$$\mathbf{E} \triangleq \begin{bmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{bmatrix}, \tag{2.77}$$

which is obtained via singular value decomposition with eigenvalues in descending order,

$$\mathbf{C}_E \triangleq \begin{bmatrix} \mathbf{E}_X^H \\ \mathbf{E}_Y^H \end{bmatrix} \begin{bmatrix} \mathbf{E}_X | \mathbf{E}_Y \end{bmatrix} = \mathbf{E} \Lambda_E \mathbf{E}^H. \tag{2.78}$$

The eigenvalues of $\mathbf{\Psi}_{TLS}$ correspond to the subarray shift, which lead to the source DOAs $\hat{\Theta}$ (defined in (2.62)). Since the array steering matrix (2.27) is frequency-dependent, a focusing matrix is usually needed to obtain the wideband covariance matrix from those of each frequency band [87, 90, 96].

For uniform circular microphone arrays (e.g. UCAs), the EB-ESPRIT [40, 97] uses the eigenbeam decomposition to achieve wideband rotational invariance. Instead of using the covariance matrix of raw signals as in (2.64) directly, the covariance matrix of circular harmonics is

$$\mathbf{R}_{EB}(\omega) = \mathbf{C}_{\mathrm{CH}}(\omega) \mathbf{C}_{\mathrm{CH}}^H(\omega), \tag{2.79}$$

where $\mathbf{C}_{\mathrm{CH}}(\omega)$ is given in (2.55). In the spherical harmonics domain, the steering array manifold vectors, i.e. (2.56), are frequency-independent. Thus the wideband covariance matrix can be simply obtained by averaging those of all frequency bins [40, 97–99]. Further steps of the EB-ESPRIT processing follows the standard ESPRIT. Since the standard MUSIC and ESPRIT methods are derived from the anechoic model, the location estimates may deviate from the true DOAs in reverberant conditions.

2.5.3 TDOA-based Methods

Using the far-field assumption and choosing the center of gravity of the microphone array as the origin of the coordinate system, from (2.21) the TDOA

from a source q to any two microphones indexed i and j is

$$\tau_{ij}(\theta_q) = \frac{(\overrightarrow{m}_i - \overrightarrow{m}_j) \cdot \overrightarrow{\wp}_q(\theta_q)}{v \, \|\overrightarrow{\wp}_q(\theta_q)\|}, \tag{2.80}$$

where the numerator is a dot product. Thus using the microphone array geometry (e.g. $\overrightarrow{m}_i$ and $\overrightarrow{m}_j$), the DOA θ_q can be found from the TDOA estimate. Note that using only two microphones or a linear array leads to the ambiguity that a TDOA may correspond to two DOAs.

GCC AND GCC-PHAT

The classical generalized cross-correlation (GCC) method uses the cross-power spectral density (CSD) function as defined in (2.5). It uses the anechoic signal model in (2.22). Compared with the standard cross-correlation function, it improves the performance of the time delay estimation by prefiltering signals prior to the cross correlation [35, 100].

Following the definitions in (2.3) and (2.5), the CSD [101] between observed signals is $\mathbb{E}[X_i(\Omega) \cdot X_j^\star(\Omega)]$. Thus using the STFT denotations (cf. (2.37)) and the WSS assumption,

$$G_{x_i x_j}(\omega) = \mathbb{E}[X_i(\omega) \cdot X_j^\star(\omega)], \tag{2.81}$$

where $[\cdot]^\star$ is the complex conjugate operation, and the expectation is usually approximated as the average over time frames within a short time interval, e.g. $\hat{G}_{x_i x_j}(\omega) = \frac{1}{N} \sum_{k=1}^{N} X_i(k,\omega) \cdot X_j^\star(k,\omega)$.

The CSD between filtered outputs is then

$$\varepsilon_{ij}^{\mathrm{gcc}}(\omega) = \Psi_{ij}(\omega) G_{x_i x_j}(\omega), \tag{2.82}$$

where $\Psi_{ij}(\omega)$ is the overall frequency response of the prefilters used. The goal is to find the time delays that corresponds to cross-correlation peaks.

Using the relationship between cross-correlation and CSD (from (2.3) and (2.5)), the cross-correlation between two prefiltered microphone signals

is

$$\epsilon_{ij}^{\text{gcc}}(\tau_{ij}(\theta)) = \sum_{\omega} \varepsilon_{ij}^{\text{gcc}}(\omega) \cdot e^{J\omega f_s \tau_{ij}(\theta)}. \tag{2.83}$$

Thus the TDOA estimate is

$$\tau_{ij}^{\text{gcc}}(\hat{\theta}) = \arg\max_{\tau_{ij}} \ \epsilon_{ij}^{\text{gcc}}(\tau_{ij}(\theta)). \tag{2.84}$$

DOA estimates $\hat{\Theta}$ (defined in (2.62)) can then be found from the peaks. $\Psi_{ij}(\omega)$ can have various forms such as Roth, SCOT, or PHAT [35]. Particularly, when $\Psi_{ij}(\omega) = 1/|G_{x_i x_j}(\omega)|$ in (2.82), (2.84) becomes the GCC-PHAT estimation $\tau_{ij}^{\text{gcc-phat}}(\hat{\theta})$ [35]. GCC methods relies on the anechoic model [35, 36], and the PHAT is most often used in practice for better performance compared to the other prefilters.

SRP-PHAT

The SRP-PHAT [43,44] can be classified either as a filter-and-sum beamforming technique, or a TDOA-based method built upon the GCC-PHAT.

Apparently, the time delay in (2.84) depends on the DOA θ. In the implementation, the localization function of SRP-PHAT is defined as:

$$\epsilon^{\text{srp-phat}}(\theta) = \sum_{i}^{I_M} \sum_{j}^{I_M} \epsilon_{ij}^{\text{gcc-phat}}(\tau_{ij}(\theta)), \tag{2.85}$$

where $\epsilon_{ij}^{\text{gcc-phat}}(\tau_{ij}(\theta))$ is defined in (2.83), using the PHAT prefilter. DOA estimates $\hat{\Theta}$ (defined in (2.62)) can then be found from the peaks.

The SRP-PHAT uses all combinatory pairs of microphones of the array, which contain redundant information. Although this improves the robustness against reverberation, the summation in (2.85) may not guarantee high location resolution. Test results are shown in Section 3.7 and Appendix F .

MCCC

The standard MCCC method developed in [42, 102] generalizes the classical cross-correlation coefficient to the multichannel case in the time domain. The far-field and anechoic assumptions are applied so that when perfectly time-aligned in each channel, the impinging signal at any microphone can be expressed as a linear combination of signals at the rest microphones. Its spatial correlation matrix for each scanned location is defined as

$$\tilde{\mathbf{R}}(\overrightarrow{\wp}(\theta)) = \mathbb{E}[\mathbf{x}_{\overrightarrow{\wp}}\mathbf{x}_{\overrightarrow{\wp}}^{T}] \triangleq \begin{bmatrix} \sigma_1^2 & r_{12} & \dots & r_{1I} \\ r_{21} & \sigma_2^2 & \dots & r_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ r_{I1} & r_{I2} & \dots & \sigma_{I_M}^2 \end{bmatrix}_{\overrightarrow{\wp}}, \tag{2.86}$$

where in $\mathbf{x}_{\overrightarrow{\wp}}$, signals are aligned in time between microphones according to the scanned location $\overrightarrow{\wp}(\theta)$. Respectively, σ_i and r_{ij} are the autocorrelation and cross-correlation coefficients. Thus when signals from microphones are perfectly aligned in $\mathbf{x}_{\overrightarrow{\wp}}$, the determinant of the spatial correlation matrix is zero, indicating that the sound source comes from location $\overrightarrow{\wp}(\theta)$.

Using the far-field model, the DOA estimate is

$$\hat{\theta} = \arg\max_{\theta} \ \rho^2_{1:I_M}(\overrightarrow{\wp}(\theta)) = \arg\min_{\theta} \det[\tilde{\mathbf{R}}(\overrightarrow{\wp}(\theta))], \tag{2.87}$$

where the localization function

$$\rho^2_{1:I_M}(\overrightarrow{\wp}(\theta)) \triangleq 1 - \frac{\det[\tilde{\mathbf{R}}(\overrightarrow{\wp}(\theta))]}{\prod_{i=1}^{I_M}\sigma_i^2}, \tag{2.88}$$

the determinant of a matrix is defined as $\det[\cdot]$. DOA estimates $\hat{\Theta}$ (defined in (2.62)) can then be found from the peaks of the localization function.

Since the integer sample shifting may not perfectly align microphone signals in time (considering the finite sampling frequency), and that the localization function relies on the determinant of the spatial matrix which consists of correlation coefficients of all microphone pairs including those far apart

(spatially aliased), the MCCC by itself may not provide good spatial resolution required for multi-speaker localization (cf. Appendix E).

NEURO-FUZZY

Human listeners have the remarkable and often unparalleled capability in accurately locating sound sources even in highly reverberant environments. Computational Auditory Scene Analysis (CASA) approaches can be essentially regarded as "machine learning" methods that model the human auditory system based on the inference of psychoacoustic studies. They have received much attention and achieved considerable useful results for the audio signal processing arena in the past few decades [39, 46, 54, 103–106].

In [39], a Neuro-Fuzzy speaker localization method was developed, based on a range of CASA techniques, namely the gammatone filterbank, ERBS (equivalent rectangular bandwidth scale), glimpsing, precedence effect, auditory nerve spikes generation, inter-aural cross-correlation, phase-locking, etc. The Neuro-Fuzzy method can localize multiple speakers in highly reverberant environments as demonstrated in [39], but it is not based on mathematical derivations and motivations of some parameters have been unclear.

2.5.4 Summary and Critiques

The location estimators as described aim to provide estimates of locations (e.g. $\hat{\Theta}$ in (2.62)). A brief summary of existing localization methods is made as follows. Note that the intention of the critiques is not to refute the values of respective localization methods, but rather to examine the underlying assumptions and difficulties in the studied challenges, and thus motivate new solutions in Chapter 3 for such cases.

1. Moving sources may not be an essential problem to beamformers, as the beamformer response, e.g. (2.40), is independent with the observed signals. However, due to the sidelobe responses, the DOA estimation can be easily offset when there are concurrent sources in other directions, or besmeared when there is considerable reverberation. Moreover, due to the beamwidth of wideband beamformers, there

can be ambiguities when concurrent sources locate in close locations, where a high resolution estimator is necessary.

2. Subspace methods can achieve good spatial resolution when the reverberation is not strong and the number of sensors is larger than the number of sources. However, the benchmark methods as discussed in Section 2.5.2 rely on the covariance matrix (2.64), which may not converge to the true values during a short period of time. For moving sources in particular, the array steering matrix (2.26) or (2.35) is time varying, which makes it more difficult for the resulting covariance matrix to converge to the true values in a short measurement time. This cannot be solved by averaging the covariance matrix over a long time, because the covariance matrix over a long observation time for moving sources is actually linearly averaging the short-time observations from different locations (cf. (2.65)), while the relationship between the steering matrix and the locations is nonlinear, even in the anechoic case. Reverberation further complicates the problem by introducing coherent sources due to sound reflections.

3. CASA techniques have been motivated essentially by the idea of making computers "work" like humans, hence can be classified into the "machine learning" category. The Neuro-Fuzzy method relies on the psychoacoustic inferences to mimic the localization mechanisms of the human auditory system. However, while it is reasonable to make machines imitate human capabilities, the underlying signal model and mathematical motivations were unclear in the literature.

4. TDOA-based methods can often provide better accuracy and resolution for concurrent sources, compared with the wideband beamformers, and are more applicable for localization of moving sources in contrast with the subspace methods [45]. Reverberation is a common problem to all these methods, and the ideal but challenging solution is to find the direct-path cues before further processing (cf. Fig. 2.1). Due to difficulties in finding the direct-path cues, simplified spatial character-

istics of reflections are often assumed in (2.35), thus the redundant information from multiple microphone pairs can be useful.

2.5.5 Localization Performance

The location estimators as summarized in the preceding part of Section 2.5 aim to produce an estimate of the location of each speaker, in each time frame. Denote the speaker location as θ, and the estimate as $\hat{\theta}$ ($\hat{\theta} \in \mathbb{R}$).

For a static speaker at θ, the distribution of the estimate $\hat{\theta}$ can be comparatively easy to characterize. The probability density function (PDF) of a random variable is completely characterized by its first- and second-order moments, if it follows the Gaussian distribution, i.e.

$$p(\hat{\theta}) = \frac{1}{\sqrt{2\pi}|\sigma_{\hat{\theta}}|} e^{-(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2/2\sigma_{\hat{\theta}}^2} \sim \mathcal{N}(\mathbb{E}[\hat{\theta}], \sigma_{\hat{\theta}}^2), \tag{2.89}$$

where the mean is defined as

$$\mathbb{E}[\hat{\theta}] \triangleq \int \hat{\theta} \cdot p(\hat{\theta}) d\hat{\theta}, \tag{2.90}$$

which is unbiased if $\mathbb{E}[\hat{\theta}] = \theta$. The variance is defined as

$$\sigma_{\hat{\theta}}^2 \triangleq \mathrm{var}(\hat{\theta}) \triangleq \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]. \tag{2.91}$$

The mean-square error (MSE) is defined as

$$\mathrm{MSE}(\hat{\theta}) \triangleq \mathbb{E}[(\hat{\theta} - \theta)^2], \tag{2.92}$$

which is also the variance, for an unbiased estimator.

The root-mean-square error (RMSE) is the positive square root of the MSE, and will be used for evaluating the performance of location estimators for static speakers.

For time-varying speaker locations, especially when the number of speakers is also time-varying, the OSPA (optimal sub-pattern assignment) metric, commonly used for measuring the performance of tracking filters, is more

appropriate and its definition will be provided in Section 2.6.6.

2.5.6 Localization for Tracking

Provided with location estimates (with errors) as the observation[6], tracking speakers (with time-varying states) using the obtained observation is another significant challenge.

Recall Fig. 2.2 in Section 2.1, the most common approach to such tracking problem is to model the stochastic process as a discrete time (frame) Markov process, which is characterized by its initial state PDF and state transition PDF. The state is not directly observed, but estimated from realizations of another related process, i.e. estimates from the localization step. The problem hence becomes the state estimation of a time-varying dynamical system.

In mathematical terms, suppose at time (frame) k, there are N_k speakers with states S_k ($\mathrm{S}_k \in \mathcal{S}$ assuming continuous state space), and M_k measurements Z_k.[7] Denote $Z_{0:l} \triangleq (Z_0, Z_1, \cdots, Z_l)$. The objective of state estimation is to find speaker states $\hat{\mathrm{S}}_k$ based on observations, e.g. via

$$\text{the expected a posteriori (EAP) } \hat{\mathrm{S}}_k = \mathbb{E}[\mathrm{S}_k | Z_{0:l}] \text{ , or,} \tag{2.93a}$$

$$\text{the maximum a posteriori (MAP) } \hat{\mathrm{S}}_k = \arg\max_{\mathrm{S}_k} p(\mathrm{S}_k | Z_{0:l}). \tag{2.93b}$$

Particularly, the state estimation is called "prediction" when $l < k$; "filtering" when $l = k$; and "smoothing" when $l > k$. The prediction and filtering are pertinent to the online estimation of system states. Therefore in what follows, the thesis focuses on the prediction and filtering. The related state estimation algorithm is called a (tracking) filter.

[6] In the state estimation paradigm, observation and measurement are used interchangeably, and tracking is often treated as a synonym of state estimation or filtering.

[7] The symbol Z is used for measurements in general. It can be the DOA estimates $\hat{\Theta}$ from the localization, or Cartesian coordinates.

2.6 Tracking Overview

Significant efforts have been made in the area of state filtering, out of which the recursive Bayesian approach [107] provides a general framework based on the discrete-time Markov model. The celebrated Kalman filter [108] can be regarded as a special case for the linear Gauss-Markov dynamical and measurement process, and provides a close-form tractable solution with minimum mean-square errors (MMSE) optimality [18]. For nonlinear state transition models, the extended Kalman filter (EKF) [109] is a first-order approximation based on local linearization, while the unscented Kalman filter (UKF) [12] propagates the first and second moments using the sampling principles of the unscented transform. The extensively used Sequential Monte Carlo (SMC) filters [110, 111] use point-mass approximations to the state PDF, and the ubiquitous particle filters perform SMC filtering via sequential importance sampling.

For uncertain measurements, data association is needed before applying the Kalman filter. Typical data association techniques include the joint probabilistic data association (JPDA) [13] and the multiple hypothesis tracking (MHT) filter [112], as well as their variants. Traditionally, the JPDA filter assumes that the number of objects is fixed and known and evaluates and gates the joint association probability of all measurements to objects. Its computational complexity grows exponentially with the total number of targets and measurements. The MHT exhaustively searches all previous time steps for all possible combinations of measurement to object associations and maintains hypotheses of high posterior probability. This way the number of hypothesis increases exponentially with time. Thus various extensions have been developed [113, 114]. However, as also summarized and articulated in [115], the Bayesian consistency and the optimality of these traditional techniques was unclear, while the Byes random finite set (RFS) approach [18, 24] gives a mathematically consistent formulation for multi-object filtering.

The basic Bayes recursion, Kalman filter, Bayes RFS and the GLMB filter are summarized as follows.

2.6.1 Bayes Recursion

From (2.93), the *a posteriori* PDF is required for the EAP or MAP state estimate. This is usually propagated by the Bayes recursion. Assume a first-order Markov process, and denote the state transition PDF from time k to $k+1$ as $f_{k+1|k}(\mathrm{S}|\mathrm{S}')$, where S and S$'$ denote state variables at time $k+1$ and k respectively. From measurement $Z_{0:k}$, the PDF of the predicted state at $k+1$ can be found using the total probability theorem, i.e.

$$p_{k+1|k}(\mathrm{S}|Z) = \int f_{k+1|k}(\mathrm{S}|\mathrm{S}')\ p_{k|k}(\mathrm{S}'|Z)\ d\mathrm{S}'\ , \tag{2.94}$$

where $p_{k+1|k}(\mathrm{S}|Z) \triangleq p(\mathrm{S}_{k+1}|Z_{0:k})$, and the prior PDF $p_{k|k}(\mathrm{S}'|Z) \triangleq p(\mathrm{S}'_k|Z_{0:k})$.

Using the Bayes' rule, the posterior PDF of the state at time $k+1$ given new measurement z_{k+1} is thus

$$p_{k+1|k+1}(\mathrm{S}|Z) = \frac{g_{k+1}(Z|\mathrm{S})\ p_{k+1|k}(\mathrm{S}|Z)}{\int g_{k+1}(Z|\mathrm{S}')\ p_{k+1|k}(\mathrm{S}'|Z)\ d\mathrm{S}'}\ , \tag{2.95}$$

where $p_{k+1|k+1}(\mathrm{S}|Z) \triangleq p(\mathrm{S}_{k+1}|Z_{0:k+1})$, and $g_{k+1}(Z|\mathrm{S}) \triangleq g(Z_{k+1}|\mathrm{S}_{k+1})$ denotes the measurement likelihood due to the *a priori* knowledge (e.g. from the localization stage in Part I). The normalizing denominator is required.

(2.94) and (2.95) form the basic Bayes prediction and filtering recursion. Depending on the state transition and measurement models, i.e. $f_{k+1|k}(\mathrm{S}|\mathrm{S}')$ and $g_{k+1}(Z|\mathrm{S})$, a good variety of state filters can be found in the literature. Particularly, the Kalman filter provides a closed-form solution to the linear "Gauss-Markov"[8] case when the prior state and noises are uncorrelated and Gaussian, and the state transition and measurement models are linear.

[8]The Gauss-Markov process model shall not be confused with the Gauss-Markov estimate.

2.6.2 Linear Gauss-Markov Model and Kalman Filter

The simple linear Gauss-Markov model in the vector form specifies [116]

$$\mathrm{S}_{k+1} = \mathcal{A}_{k+1}\mathrm{S}_k + \mathcal{B}_{k+1}\mathbf{u}_{k+1} \tag{2.96a}$$

$$Z_{k+1} = \mathcal{H}_{k+1}\mathrm{S}_{k+1} + \mathbf{w}_{k+1}, \tag{2.96b}$$

where $\mathcal{A}_{k+1}$ (state transition), $\mathcal{B}_{k+1}$, and $\mathcal{H}_{k+1}$ (observation) are *a priori* known *matrices*, $\mathbf{u}_{k+1} \sim \mathcal{N}(\mathbf{0}, \Sigma_u)$ the driving noise vector, and $\mathbf{w}_{k+1} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathrm{w})$ the observation noise vector.

The Kalman filter provides the optimal MMSE recursive solution [116]

$$\text{Prediction} \quad \hat{\mathrm{S}}_{k+1|k} = \mathcal{A}_{k+1}\mathrm{S}_k \tag{2.97a}$$

$$\text{Prediction MMSE} \quad \mathcal{M}_{k+1|k} = \mathcal{A}_{k+1}\mathcal{M}_k\mathcal{A}_{k+1}^T + \mathcal{B}_{k+1}\Sigma_u\mathcal{B}_{k+1}^T \tag{2.97b}$$

$$\text{Kalman Gain} \quad \mathcal{K}_{k+1} = \mathcal{M}_{k+1|k}\mathcal{H}_{k+1}^T(\Sigma_\mathrm{w} + \mathcal{H}_{k+1}^T\mathcal{M}_{k+1|k}\mathcal{H}_{k+1}^T)^{-1} \tag{2.97c}$$

$$\text{Correction} \quad \hat{\mathrm{S}}_{k+1} = \hat{\mathrm{S}}_{k+1|k} + \mathcal{K}_{k+1}(\mathcal{Z}_{k+1} - \mathcal{H}_{k+1}\hat{\mathrm{S}}_{k+1|k}) \tag{2.97d}$$

$$\text{MMSE Matrix} \quad \mathcal{M}_{k+1} = (\mathbf{I} - \mathcal{K}_{k+1}\mathcal{H}_{k+1})\mathcal{M}_{k+1|k}. \tag{2.97e}$$

For more general scenarios, e.g. nonlinear state transition and measurement models, the EKF and UKF can be used, but the linearization preserves no optimality. However, the finite set statistics (FISST) theory provides a neat solution in such cases [24], representing the measurements and states with random finite sets (RFS's) and retaining the Bayesian recursion formalism.

2.6.3 FISST and Bayes RFS Filters

In practice, the measurements (e.g. multi-speaker location estimates $\hat{\Theta}$) are often sets of finite *unordered* estimates, indicating that the association is *unknown and time-varying* between state estimates and measurements.

Denote the RFS multi-target posterior density at time k as $\pi_k(\cdot|Z)$. Similar to (2.94) and (2.95), it is propagated by the multi-object Bayes recursion

(see e.g. [21] or [18, Eq.(14.14), (14.50)]):

$$\pi_{k+1|k}(\mathrm{S}|Z) = \int f_{k+1|k}(\mathrm{S}|\mathrm{S}')\pi_k(\mathrm{S}'|Z)\delta\mathrm{S}' \,, \tag{2.98a}$$

$$\pi_{k+1}(\mathrm{S}|Z) = \frac{g_{k+1}(Z|\mathrm{S})\pi_{k+1|k}(\mathrm{S}|Z)}{\int g_{k+1}(Z|\mathrm{S}')\pi_{k+1|k}(\mathrm{S}'|Z)\delta\mathrm{S}'} \,, \tag{2.98b}$$

where $\pi_{k+1|k}(\mathrm{S}|Z) \triangleq p(\mathrm{S}_{k+1}|Z_{0:k})$, $\pi_k(\mathrm{S}|Z) \triangleq p(\mathrm{S}_k|Z_{0:k})$, and $\pi_{k+1}(\mathrm{S}|Z) \triangleq p(\mathrm{S}_{k+1}|Z_{0:k+1})$. States S and measurements Z are now RFS's.

The standard dynamical and measurement model assumes that, each object (with state $\mathrm{s}_{i_k} \in \mathrm{S}_k$) either continues to exist at time $k+1$ with probability $p_{\mathrm{S},k+1}(\mathrm{s}_{i_k})$ and transit to a new state $\mathrm{s}_{i_{k+1}}$ with probability density $f_{k+1|k}(\mathrm{s}_{i_{k+1}}|\mathrm{s}_{i_k})$, or dies with probability $1 - p_{\mathrm{S},k+1}(\mathrm{s}_{i_k})$. Meanwhile, s_{i_k} is either detected with probability $p_{D,k}(\mathrm{s}_{i_k})$ and generates an measurement z_k with likelihood $g_k(z_k|\mathrm{s}_{i_k})$, or missed with probability $1 - p_{D,k}(\mathrm{s}_{i_k})$. In addition, due to newborn targets, there are spontaneous births that constitute S_{k+1}, and due to non-ideal sensors, there are false alarms or clutter with intensity $\kappa_k(\cdot)$. These parameters will be detailed later. For notational simplicity, the functional dependence on time indices are omitted in the following.

The integrals in (2.98a) and (2.98b) are FISST (finite set statistics) set integrals [18], which can be intractable. Practical implementations usually require approximations (e.g. (C)PHD [16,19], CBMB [20], (G)LMB [21,22]) to make the recursion tractable without severely degrading the accuracy.

The thesis focuses on the application of the state-of-the-art (G)LMB filter, which not only provides estimations of the multi-object states, but also jointly tracks the associations (identities) of states over time for each object. This is an important advancement for multi-speaker tracking. For completeness, a brief summary of the (G)LMB filter is provided as follows.

2.6.4 GLMB RFS Filter

GLMB RFS

The GLMB RFS $\mathbf{S} \triangleq \{\mathbf{s}_i = (\mathrm{s}_i, \ell_i) \mid i \in \mathbb{N}\}$ is a closed-form solution to the multi-object Bayes RFS recursion. It is a labeled RFS with state space $\mathbb{S}$ and

label space $\mathbb{L}, (\ell_i \in \mathbb{L})$, where the labels are unique, i.e. $\ell_i \neq \ell_{i'}, \forall i \neq i'$. The GLMB distribution is written as

$$\mathfrak{B}(\mathbf{S}) = \Delta(\mathbf{S}) \sum_{\xi \in \Xi} w^{(\xi)}(\mathcal{L}(\mathbf{S})) \left[p^{(\xi)}\right]^{\mathbf{S}}, \tag{2.99}$$

where $\Delta(\mathbf{S})$ is the distinct label indicator. $p^{(\xi)}$ is the probability distribution of a target state, ξ represents a history of association map between targets and measurements, Ξ is a discrete space, $w^{(\xi)}(L)$ is the probability of hypothesis, and the multi-object exponential $\left[p^{(\xi)}\right]^{\mathbf{S}} \triangleq \prod_{\mathbf{s} \in \mathbf{S}} \left[p^{(\xi)}\right]^{\mathbf{s}}$. The projection $\mathcal{L}((\mathrm{s}, \ell)) = \ell$, and $\mathcal{L}(\mathbf{S}) = \{\mathcal{L}(\mathbf{s}) : \mathbf{s} \in \mathbf{S}\}$. $\sum_{L \in \mathbb{L}} \sum_{\xi \in \Xi} w^{(\xi)}(L) = 1$, where L is the set of labels.

Based on the GLMB distributions, the alternative δ-GLMB form is provided to facilitate numerical implementation [21]

$$\mathfrak{B}(\mathbf{S}) = \Delta(\mathbf{S}) \sum_{(L,\xi) \in \mathcal{F}(\mathbb{L}) \times \Xi} \omega^{(L,\xi)} \delta_L(\mathcal{L}(\mathbf{S})) \left[p^{(\xi)}\right]^{\mathbf{S}}, \tag{2.100}$$

where $\omega^{(L,\xi)} = w^{(\xi)}(L)$ is the probability of the hypothesis (L, ξ). $\delta_L(\mathcal{L}(\mathbf{S}))$ is the generalized Kronecker delta function for RFS denotations, which indicates whether the set of labels in $\mathbf{S}$ matches that of L. The δ-GLMB is completely characterized by the set of parameters $\{(\omega^{(L,\xi)}, p^{(\xi)}) : (L, \xi) \in \mathcal{F}(\mathbb{L}) \times \Xi\}$. (See e.g. [21–23] and the references therein for detailed studies of the (G)LMB and δ-GLMB RFS tracking filters.)

Similar to (2.98a) and (2.98b), the GLMB recursion consists of the multi-object "update" step based on Bayes inference and the Chapman-Kolmogorov [117] "prediction" step based on the state dynamical models.

UPDATE

If the current RFS prediction density is a δ-GLMB of the form (2.100), using the current multi-feature measurement Z, the posterior density is still a δ-

GLMB [22], i.e.

$$\text{ß}(\mathbf{S}|Z) = \Delta(\mathbf{S}) \sum_{(L,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \sum_{\vartheta\in\boldsymbol{\Theta}(L)} \omega^{(L,\xi,\vartheta)}(Z)\delta_L(\mathcal{L}(\mathbf{S}))\left[p^{(\xi,\vartheta)}(\cdot|Z)\right]^{\mathbf{S}}, \tag{2.101}$$

where for notational convenience, the dependence on k is suppressed,

$$\omega^{(L,\xi,\vartheta)}(Z) \propto \omega^{(L,\xi)}[\eta_Z^{(\xi,\vartheta)}]^L \tag{2.102a}$$

$$\eta_Z^{(\xi,\vartheta)}(\ell) = \left\langle p^{(\xi)}(\cdot,\ell), \psi_Z(\cdot,\ell;\vartheta)\right\rangle \tag{2.102b}$$

$$p^{(\xi,\vartheta)}(\mathrm{s},\ell|Z) = \frac{p^{(\xi)}(\mathrm{s},\ell)\psi_Z(\mathrm{s},\ell;\vartheta)}{\eta_Z^{(\xi,\vartheta)}(\ell)} \tag{2.102c}$$

$$\psi_Z(\mathrm{s},\ell;\vartheta) = \begin{cases} \frac{p_D(\mathrm{s},\ell)g(z_{\vartheta(\ell)}|\mathrm{s},\ell)}{\kappa(z_{\vartheta(\ell)})}, \text{if } \vartheta(\ell) > 0 \\ 1 - p_D(\mathrm{s},\ell), \text{if } \vartheta(\ell) = 0\,, \end{cases} \tag{2.102d}$$

$\langle\cdot,\cdot\rangle$ denotes inner product, $g(z_{\vartheta(\ell)}|\mathrm{s},\ell)$ is the single object likelihood for the measurement $z_{\vartheta(\ell)}$ being generated by (s,ℓ), and $\kappa(\cdot)$ is the intensity function of Poisson RFS describing the clutter. $\boldsymbol{\Theta}(L)$ denotes the subset of current association maps with domain L.

After update, the maximum *a posteriori* (MAP) estimate of the cardinality (number of speakers) is chosen, and the corresponding hypothesis with the highest weight is used for the multi-object state estimate.

PREDICTION

If the current RFS filtering density from its previous update step is a δ-GLMB of the form (2.100), the prediction density to the next time is also a δ-GLMB given as [22]

$$\text{ß}_+(\mathbf{S}_+) = \Delta(\mathbf{S}_+) \sum_{(L_+,\xi)\in\mathcal{F}(\mathbb{L}_+)\times\Xi} \omega_+^{(L_+,\xi)}\delta_{L_+}(\mathcal{L}(\mathbf{S}_+))\left[p_+^{(\xi)}\right]^{\mathbf{S}_+}, \tag{2.103}$$

where $[\cdot]_+$ stands for prediction.

$$\omega_+^{(L_+,\xi)} = \omega_S^{(\xi)}(L_+\cap\mathbb{L})w_B(L_+\cap\mathbb{B}) \tag{2.104a}$$

$$\omega_S^{(\xi)}(L) = [\eta_S^{(\xi)}]^L \sum_{I \supseteq L} [1 - \eta_S^{(\xi)}]^{I-L} \omega^{(L,\xi)} \tag{2.104b}$$

$$\eta_S^{(\xi)}(\ell) = \left\langle p_S(\cdot, \ell), p^{(\xi)}(\cdot, \ell) \right\rangle \tag{2.104c}$$

$$p_+^{(\xi)}(\mathrm{s}, \ell) = 1_{\mathbb{L}}(\ell) p_S^{(\xi)}(\mathrm{s}, \ell) + 1_{\mathbb{B}}(\ell) p_B(\mathrm{s}, \ell) \tag{2.104d}$$

$$p_S^{(\xi)}(\mathrm{s}, \ell) = \frac{\left\langle p_S(\cdot, \ell) f(\mathrm{s}|\cdot, \ell), p^{(\xi)}(\cdot, \ell) \right\rangle}{\eta_S^{(\xi)}(\ell)} \tag{2.104e}$$

The inclusion function, a generalization of the indicator function, is defined by

$$1_{\mathscr{Y}}(\mathscr{X}) \triangleq \begin{cases} 1, \text{ if } \mathscr{X} \subseteq \mathscr{Y} \\ 0, \text{ otherwise.} \end{cases} \tag{2.105}$$

$\mathbb{B}$ is the space of newborn target labels. The set of newborn targets can be represented by an LMB RFS $\{(w_B, p_B)\}$, where w_B is the probability of a birth hypothesis of newborn targets and p_B is the probability distribution of kinematic states that belong to the birth targets.

Since the form of the GLMB PDF is retained in the recursion, it is called a conjugate prior [21]. The standard implementation of GLMB filter assumes known birth probability densities $\{(w_B, p_B)\}$, which can be restricting in practice. Moreover, the standard implementation of the (G)LMB filter have been focusing on the single-feature state (e.g. kinematic location and speed) filtering, which may be subject to the ambiguity problem when the states of different objects are too close. Although the expressions of the Bayes RFS filters may appear intricate, they are indeed neat *closed-form* solutions to multi-object tracking, and can be computationally efficient [118, 119].

MDB Model

Standard implementations of the GLMB filters require *a priori* knowledge of object birth distributions, and therefore can be restrictive in practical applications. An adaptive birth model for Sequential Monte Carlo (SMC) implementations of PHD and CPHD filters has been proposed in [120]. An MDB for SMC-CBMeMBer has been presented in [121]. The adaptive birth distribution for the LMB filter has also been proposed [23]. Similarly, [25] provides

the MDB model for the GLMB filter described as follows.

The MDB model adaptively initiates the kinematic states and existence probabilities of birth objects based on measurement data from previous time, thereby eliminating the dependence of *a priori* knowledge of object birth distributions.

Suppose measurements Z are not associated with any persistent object labels at the current time frame. At the next time step, these measurements will then initiate new-born objects. The set of new-born objects can be completely characterized by an LMB RFS, i.e. $\{r_B^{(\ell)}(z),\ p_B(\cdot,\cdot;z) : \ell = \ell_B(z)\}_{z\in Z}$ where $r_B(z)$ denotes the existence probability of the non-empty birth object initiated by measurement z, $\ell_B(z)$ denotes the assigned label, and $p_B(\mathrm{s},\ell;z)$ is the probability density of the corresponding birth object.

The probability density of the new-born LMB RFS is thus

$$\mathfrak{B}_B(\mathbf{S}_+) = \Delta(\mathbf{S}_+)w_B(\mathcal{L}(\mathbf{S}_+))\,[p_B]^{\mathbf{S}_+}, \tag{2.106}$$

where as used in (2.104a),

$$w_B(L) = \prod_{i\in\mathbb{B}}\left(1-r_B^{(i)}\right)\prod_{\ell\in L}\frac{1_{\mathbb{B}}(\ell)r_B^{(\ell)}}{1-r_B^{(\ell)}}. \tag{2.107}$$

The new-born likelihood for each measurement $z \in Z$ can be found by

$$r_U(z) = 1 - \sum_{(L,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi}\ \sum_{\vartheta\in\boldsymbol{\Theta}(L)} 1_{z_\vartheta}(z)\omega^{(L,\xi,\vartheta)}, \tag{2.108}$$

where $\omega^{(L,\xi,\vartheta)}$ is given in (2.102a). It can be seen from (2.108) that, $r_U(z) = 0$ if a measurement has been used in all hypotheses, while $r_U(z) = 1$ for measurements that have not been assigned to any of the objects.

In (2.107), the existence probability of the new-born object is obtained via

$$r_B(z) = \min\left(r_{B_{\max}},\ \lambda_B\cdot\frac{r_U(z)}{\sum_{\zeta\in Z}r_U(\zeta)}\right), \tag{2.109}$$

where λ_B is the expected number of object birth at the next time, and $r_{B_{\max}} \in [0,1]$ is the maximum existence probability of a new-born object.

The mean cardinality of the new-born labeled multi-Bernoulli RFS is

$$\sum_{\zeta \in Z} r_B(\zeta) \leq \lambda_B. \tag{2.110}$$

A new birth of Bernoulli RFS is generated around the measurement, for each measurement z that has non-zero new-born likelihood. Assuming a Gaussian distribution, the probability distribution of the states is given in (2.111), which is used in (2.104d) of the GLMB filter.

$$p_B(\mathrm{s}, \ell; z) = \sum_{i=1}^{M_b} \frac{1}{M_b} \delta_{\mathrm{s}_z^{(i)}}(\mathrm{s}),\ z \in Z, \tag{2.111}$$

$$\mathrm{s}_z^{(i)} \sim \mathcal{N}\big(\mathrm{s}; m_B(z), P_B(z)\big),\ i = 1, ..., M_b, \tag{2.112}$$

where M_b denotes the number of generated states, $m_B(z)$ a function mapping from an observation to its corresponding object state, and $P_B(z)$ the variance.

2.6.5 Multi-speaker Tracking

Although it may appear straightforward in using the MDB GLMB based filters for the adaptive multi-speaker tracking (or feature filtering in general), the benefits and challenges to be addressed are summarized as follows.

- The GLMB filter provides a closed-form solution, tracking not only the kinematic states, but also the speaker identities (labels) jointly. Kinematic states can be expressed in terms of DOAs or Cartesian coordinates (based on the triangulation using multiple microphone arrays). Considerations in converting the DOAs to Cartesian coordinates, especially when using the MDB model, will be detailed in Part II.

- In some applications, e.g. the bearing and range tracking, certain kinematic state transition models are assumed fully known *a priori* [21, 22, 25]. In speaker tracking however, it is rather impractical to assume a completely known state transition function. Thus the Langevin model is often used in the context of speaker tracking, to accommodate

the random walks. Moreover, in the challenging reverberant environments, the robust localization methods are essential in order to obtain reliable tracking results. Studies on multi-speaker tracking with strong reverberation will be carried out in Part II.

- The standard implementation of the GLMB filter deals with only one kinematic state [21, 22, 25], which may not be able to resolve the ambiguity when multiple speakers locate closely, and hence more speaker features are needed. Moreover, it is of practical importance to separate other speaker features (e.g. speech signals, pitches) besides the locations. Thus jointly filtering (i.e. tracking and separating) multiple features of multiple speakers is an interesting challenge and its feasibility will be investigated in Part II.

2.6.6 Tracking Performance

The commonly used RMSE measure may suffice the simple cases such as a single speaker or static speakers, but not otherwise when the cardinality error shall be taken into consideration (e.g. for clutter and miss-detections).

OSPA Metric

The estimation accuracy of the multi-object localization and tracking methods is more closely evaluated using the OSPA (optimal sub-pattern assignment) [122] metric, as it not only evaluates the location miss-distances but also assesses the cardinality errors. The OSPA metric $\bar{d}_p^{(c)}$ of two arbitrary finite sets $\mathrm{S} = \{\mathrm{s}_1, ..., \mathrm{s}_m\}$ and $\hat{\mathrm{S}} = \{\hat{\mathrm{s}}_1, , ..., \hat{\mathrm{s}}_n\}, (0 \leq m \leq n)$ is defined as

$$\bar{d}_p^{(c)}(\mathrm{S}, \hat{\mathrm{S}}) \triangleq \Big(\frac{1}{n} \big(\min_{\pi \in \Pi_n} \sum_{i=1}^{m} d^{(c)}(\mathrm{s}_i, \hat{\mathrm{s}}_{\pi(i)})^p + c^p (n-m) \big) \Big)^{\frac{1}{p}}, \tag{2.113}$$

where $p \geq 1$, $c > 0$, $d^{(c)}(\mathrm{s}, \hat{\mathrm{s}}) \triangleq \min(c, \|\mathrm{s} - \hat{\mathrm{s}}\|)$, and Π_n denotes the set of permutations on $\{1, 2, ..., n\}$, $\forall n \in \mathbb{N}$. The distance $\bar{d}_p^{(c)}(\mathrm{S}, \hat{\mathrm{S}})$ is interpreted as a p-th order per-object error. If $m > n$, $\bar{d}_p^{(c)}(\mathrm{S}, \hat{\mathrm{S}}) = \bar{d}_p^{(c)}(\hat{\mathrm{S}}, \mathrm{S})$. The order

parameter p determines the sensitivity to outliers, and the cut-off distance c determines the weighting for cardinality errors.

2.7 Problem Formulation and Proposed System

Based on the signal and system models as well as the discussions in previous sections, the proposed solution to the studied problem is shown in Fig. 2.5. The work aims to extract locations (and sounds) of respective speakers from the microphone recordings, based on the *a priori* knowledge of microphone array geometries. The number of active speakers is unknown and time-varying, and speakers can move while talking.

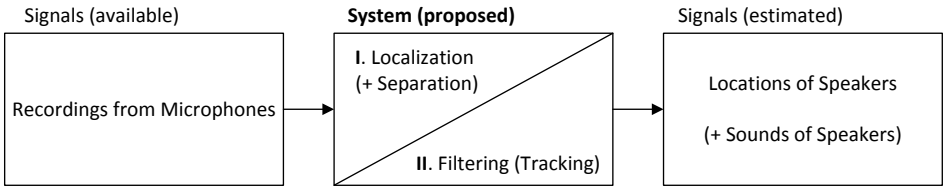

Figure 2.5: Formulation of the proposed system.

Recall Fig. 2.2, as shown in the general signal models (2.30) and (2.34), the discrete signals available are the $\mathbf{x}[n]$ and $\mathbf{X}(\omega)$, respectively. The desired signals are the speaker locations, e.g. θ_q in (2.28) (and corresponding sounds signals $s_q(t)$), which are embedded in, and to be estimated from, the available signal snapshots *in short time frames* and *over time*.

The proposed system consists of two major parts in logical order, as will be further elaborated in what follows. Briefly speaking, Part I is regarding speaker localization[9] in short time frames, and Part II is mainly the feature filtering (e.g. tracking) over time. The localization problem may be less challenging *if* there is no sound reflections, the speakers are static (the locations do not change over time), and the speakers talk for long periods of

[9]Note that the speech separation is not the main focus of this thesis. It is hence simply dealt with wideband beamforming techniques using location estimates.

time so that there is sufficient data for each speaker location. For the proposed localization methods in this thesis however, *none* of these assumptions is presumed. Therefore as developed in the sequel (cf. Fig. 2.5), Part I deals with the reverberant speaker localization in short time frames. Location (and sound signal) estimates obtained from Part I are used as "observations" for estimating time-varying speaker states in Part II.

Part I

Speaker Localization

Chapter 3

Speaker Localization

Acoustic localization of speakers has been an important audio signal processing problem, which arises from the practical requirements of speech acquisition, separation, recognition, transcription and speaker tracking. This chapter studies the speaker localization in reverberant conditions in short time frames. Some contributions of this chapter have been published in the author's journal article[1] [45]. Further studies such as the Onset-GSRP and the Onset-MCC implementation, the investigation of the redundant information, more test results and comparative studies are also included.

3.1 Introduction

Despite the significant literature on sound source localization, challenges still remain pertaining to strong reverberation and moving speakers. Section 2.5 has grouped speaker localization methods into three categories, based on beamforming, subspace and TDOA techniques. A beamformer scans DOAs and find peaks in the corresponding response. Beside traditional wideband beamformers, the circular harmonics beamformer (CHB) [32–34] has attracted recent attentions in the wideband source localization applications. Prominent examples include the TF - CHB [34], and the EB - ESPRIT [40] methods. Albeit straightforward, a wideband beamformer usually has comparatively wide beamwidth, hence limited DOA resolution versus the number of sensors. The sidelobe beampattern and concurrent sources may easily offset peak locations of the steered-response. Subspace-based methods decomposes the covariance matrix for narrowband parameter estimation. The popular MUSIC (multiple signal classification) [27], ESPRIT (estimation of signal parameters via rotational invariance techniques) [28] and their variants [29, 40, 123, 124], can provide good direction-of-arrival (DOA) resolu-

[1] ©2018 IEEE. Reverberation-Robust Localization of Speakers Using Distinct Speech Onsets and Multichannel Cross-Correlations, by Shoufeng Lin; IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 2018 Nov 1;26(11):2098-111.

tions for spatially uncorrelated narrowband sources, but still are not ideal concerning strong reverberation and moving speakers.

TDOA estimation algorithms mainly rely on the generalized cross-correlation (GCC) [35] or the eigenvalue decomposition (EVD) [36] method. The EVD method is not straightforward for multiple concurrent speakers, or for common zeros between impulse responses of speaker-microphone channels [36]. The popular GCC method, including its classic phase transform (PHAT) prefilter, assumes free-field plane wave model and is thus considered sensitive to strong reverberation. The SRP-PHAT is a reverberation-robust extension to the GCC-PHAT, but as will be shown further in this chapter, it may not be able to resolve closely located sources. The multichannel cross-correlation coefficient (MCCC) method [41, 102] was motivated to leverage the redundant information from multiple microphones. Using the determinant of the spatial correlation matrix however, does not provide good localization resolution for the MCCC, and although intuitive, it was unclear why the redundant information can improve reverberation robustness. For computationally viable localization implementations using multichannel cross-correlations, the generalized steered response power (GSRP) method [125] has been developed, by inversely mapping relative time delays to spatial locations. To address the reverberation problem, many exploited the "precedence effect" by extracting speech onsets (cf. Section 2.2.2) as the reliable localization cues [37, 38, 126–130]. However, the existing onset detection methods such as [37] and [39] are built upon psychoacoustic experimental inferences (e.g. the "glimpses" [131], "precedence effect" [132] and neural spikes generation [46, 133]), rather than mathematical motivations and derivations.

Following the discussion in Section 2.5.4, this chapter presents three algorithms for localizing speakers using microphone array recordings of reverberated sounds. The first two algorithms build upon an RIR model and use direct-path signal components for reliable localization. To suppress reverberation, speech onsets are detected and encoded to formulate the multichannel cross-correlation coefficients in each subband. For computationally viable implementation of DOA scan, the GSRP reverse mapping is used, and the resulting localization estimators are thus referred to as the Onset-GSRP and

the Onset-MCC. The third algorithm builds upon an RTF model and extends the GCC-PHAT method by using redundant information of multiple microphones to address the reverberation problem. Compared to the SRP-PHAT method, the proposed MCC-PHAT has improved spatial resolution. The proposed methods have been evaluated under adverse conditions using not only simulated signals (reverberation time T_{60} of up to 1s) but also recordings in a real reverberant room ($T_{60} \approx 0.65$s). Comparing with some state-of-the-art localization methods, experimental results confirm that the proposed methods can reliably locate static and moving speakers, in presence of strong reverberation.

The chapter is organized as follows. Section 3.2 provides time domain models for voiced sounds and the RIR. Section 3.4 proposes the Onset-GSRP and Onset-MCC algorithms, using the detected and encoded speech onsets as derived in Section 3.3. Section 3.5 exploits the redundant information and proposes the MCC-PHAT method. The notations for DOA estimates are provided in Section 3.6. Performance evaluations of proposed localization methods are demonstrated in Section 3.7. Conclusions are given in Section 3.8.

3.2 Time Domain Models

3.2.1 Voiced Speech Model

From Section 2.3, in presence of sound reflections, it is extremely challenging to find direct-path cues without much *a priori* knowledge. Thus the speech onset is often exploited. Based on the source excitation - vocal tract model (cf. (2.6)) for the process of speech production [48, 51], as well as the amplitude and frequency modulation structure [134], a harmonic model can be used for *voiced* sounds, i.e.

$$s_q(t) = \sum_{\hbar=1}^{H_q} s_q^{(\hbar)}(t), \tag{3.1}$$

where,

$$s_q^{(\hbar)}(t) = A_q^{(\hbar)}(t) \cdot \cos\left(\hbar \cdot \Omega_q \cdot t + \phi_q^{(\hbar)}(t)\right), \tag{3.2}$$

where $s_q(t)$ is the voiced speech signal from the q-th speaker, q is an integer index, $s_q^{(\hbar)}(t)$ the $\hbar$-th harmonic, integer $\hbar$ the order of harmonics, integer H_q the maximum order of harmonics, $A_q^{(\hbar)}(t) \geq 0$ the envelope of each harmonic, $\phi_q^{(\hbar)}(t) \in \mathbb{R}$ the phase (which is assumed constant for a short interval of time), $\Omega_q > 0$ the angular fundamental frequency, which is usually different for concurrent speakers. Compared to the modulating harmonic frequency, the bandwidth of $A_q^{(\hbar)}(t)$ is usually small. The speech onset corresponds to a rising ramp of $A_q^{(\hbar)}(t)$ in this model.

3.2.2 Reverberation RIR Model

Following the discussions in Section 2.3, the acoustic RIR from an arbitrary source q to sensor i can also be parametrized directly with the direct-path and reflection responses. The early reflections typically contain some discrete reflections, while the diffuse part is normally distributed with exponentially decaying envelope. Assuming that sound sources are not located too close to reflective surfaces, hence the early reflections are negligible [45, 68]. Thus the RIR model is written as

$$\mathrm{h}_{qi}(t) = \begin{cases} \mathrm{h}_{d_{qi}}, \ t = t_{d_{qi}} \\ \mathrm{h}_{d_{qi}} \cdot \nu_{qi}(t - t_{d_{qi}}) \cdot 10^{-3 \cdot \frac{t - t_{d_{qi}}}{T_{60}}}, \ t \geq t_{d_{qi}} + \tau_{qi} \\ 0, \text{ otherwise} \end{cases}, \tag{3.3}$$

where $t_{d_{qi}} > 0$ is the sound travelling time via the direct-path, $\mathrm{h}_{d_{qi}} > 0$ is the magnitude of the direct-path impulse response, $\nu_{qi}(t - t_{d_{qi}}) \sim \mathcal{N}(0,1)$ is a random variable, and $\tau_{qi} > 0$ is the time duration of the early reflections, which is also the delay for the first (usually strongest) diffuse reflection to arrive after the direct-path. The direct-path TDOA between two microphones is hence $\tau_{ij} = t_{d_{qi}} - t_{d_{qj}}$. The actual RIR is a realization of the model.

The early development of time domain RIR model is often accredited

to Polack [135, 136], which can be viewed as a special case of (3.3) when $t_{d_{qi}} = 0$ and $\tau_{qi} = 0$. Using the property of mathematical expectation $\mathbb{E}[\cdot]$, it is easy to check that $\mathbb{E}[\mathrm{h}_{qi}^2(t_{d_{qi}} + T_{60})] = \mathbb{E}[\mathrm{h}_{qi}^2(t_{d_{qi}})] \cdot 10^{-6}$, which is 60dB below the direct-path response $\mathbb{E}[\mathrm{h}_{qi}^2(t_{d_{qi}})] = \mathrm{h}_{d_{qi}}^2$.

Note that τ_{qi} in (3.3) is usually over a few milliseconds (ms) so that one can make a distinction between the direct-path and reflections. This can be connected with some psychoacoustic observations that when there is no time delay, a human listener may fuse two click sounds into a perceived "phantom" source between them, which shifts towards the leading sound as time delays increase to 1ms. With delays between 1ms and the "echo threshold" (a few ms), one may hear only the leading sound [69, 132]. Therefore, the RIR model in (3.3) assumes that the early reflections within a few ms following the direct-path are negligible.

3.3 Onset Detection

3.3.1 Subband Decomposition via Auditory Filterbank

From (2.29), the voiced signal at microphone i can be represented as

$$x_i(t) \approx \sum_{q=1}^{Q} \hat{s}_q(t) * \hat{\mathrm{h}}_{qi}(t) + n_i(t), \tag{3.4}$$

where $\hat{s}_q(t)$ stands for a realization of the voiced speech signal model (3.1), and $\hat{\mathrm{h}}_{qi}(t)$ denotes a realization of the RIR model in (3.3).

Based on the TF sparsity assumption [50] and the harmonic structure of speech signal (3.1), to separate signal components from different speakers, signals of each microphone can be decomposed spectrotemporally via an auditory filterbank so that speech components from separate speakers do not overlap much in each subband (see e.g. [37, 39, 46, 57, 58])

$$x_i^{(b)}(t) = x_i(t) * g^{(b)}(t), \tag{3.5}$$

where $x_i^{(b)}(t)$ denotes the decomposed signals from the i-th microphone in

subband b, and $g^{(b)}(t)$ is the filter impulse response of subband b, which is aligned in time between subbands. Common auditory filters include the gammatone filter [46, 57, 58], gammachirp filter, etc.

From (3.4) and (3.5), for the duration when the noise is small in the particular subband, the decomposed signal in subband b becomes:

$$x_i^{(b)}(t) \approx \sum_{q=1}^{Q} \hat{s}_q(t) * \hat{\mathrm{h}}_{qi}(t) * g^{(b)}(t). \tag{3.6}$$

Moreover, using the TF sparsity and (3.1), for the harmonic component $\hbar$ of the q-th speaker that falls within the passband of subband b, (3.6) further simplifies to

$$x_i^{(b)}(t) \approx \hat{s}_q^{(\hbar)}(t) * \hat{\mathrm{h}}_{qi}(t) * g^{(b)}(t), \tag{3.7}$$

where the linearity, commutativity and associativity properties of convolution, and the frequency selectivity of the filterbank are used.

In reverberant environments, the subband signals $x_i^{(b)}(t)$ in (3.7) are mixtures of direct-path and reflection components. Locations of speakers can be found via detecting the direct-paths and suppressing random reflections.

3.3.2 Speech Onsets, Direct-paths and Reflections

Suppose there is an arbitrary distinct speech onset from speaker q beginning at time $t_{on} \in \mathbb{R}$. From (3.3) it arrives at the microphone i via direct-path at time

$$t_{qi} = t_{on} + t_{d_{qi}}. \tag{3.8}$$

Assume that the reflections of preceding signals are negligible in comparison with a distinct onset. Using the RIR model in Section 2.3.2, it can be found that the diffuse reflections begin to arrive at $t_{qi} + \tau_{qi}$. Thus by expanding the convolution $\hat{s}_q^{(\hbar)}(t) * \hat{\mathrm{h}}_{qi}(t)$ in (3.7), at the vicinity of the distinct onset, $x_i^{(b)}(t)$ is composed of its direct-path and reflections, i.e.

$$x_i^{(b)}(t) = x_{d_i}^{(b)}(t) + x_{R_i}^{(b)}(t),\ t \geq t_{qi}, \tag{3.9}$$

where the direct-path component is

$$x_{d_i}^{(b)}(t) \triangleq [\hat{s}_q^{(\hbar)}(t - t_{d_{qi}}) \cdot \hat{\mathrm{h}}_{qi}(t_{d_{qi}})] * g^{(b)}(t),\ t \geq t_{qi}, \tag{3.10}$$

and from (3.3) the reflections are

$$\begin{aligned} x_{R_i}^{(b)}(t) \triangleq & \Big[\int_{\tau_{qi}}^{\infty} \hat{s}_q^{(\hbar)}(t - t_{d_{qi}} - \tau) \cdot \hat{\mathrm{h}}_{qi}(t_{d_{qi}} + \tau) d\tau \Big] * g^{(b)}(t) \\ = & \mathrm{h}_R(t) * x_{d_i}^{(b)}(t),\ t \geq t_{qi} + \tau_{qi}, \end{aligned} \tag{3.11}$$

where $\mathrm{h}_R(t)$ can be viewed as the impulse response:

$$\mathrm{h}_R(t) = \begin{cases} 0,\ t < \tau_{qi} \\ \hat{v}_{qi}(t) \cdot 10^{-3\frac{t}{T_{60}}},\ t \geq \tau_{qi}, \end{cases} \tag{3.12}$$

which represents a linear time-invariant (LTI) system, connecting an arbitrary direct-path signal and its random reflections, for a distinct onset.

3.3.3 Upper Bound of Reflection Level

It can be seen from (3.12) that the exact values of reflections are unknown without the complete knowledge of $\hat{\mathrm{h}}_{qi}(t)$, especially the $\hat{v}_{qi}(t)$ term. Thus using the property that $\mathbb{E}(|\nu_{qi}(t)|) \equiv 1$, an upper bound of the level of reflections can be used instead, which is independent on $\hat{v}_{qi}(t)$.

Using (3.11) and from Appendix C , the level of reflections is

$$\mathbb{E}\big(\lfloor x_{R_i}^{(b)}(t) \rfloor\big) \leq \tilde{\mathrm{h}}_R(t) * \lfloor x_{d_i}^{(b)}(t) \rfloor \triangleq \tilde{x}_{R_i}^{(b)}(t), \tag{3.13}$$

where $\lfloor \cdot \rfloor$ is the half-wave rectification commonly used [137–139], i.e. $\lfloor x \rfloor = \frac{1}{2}(x + |x|), \forall x \in \mathbb{R}$. $\tilde{x}_{R_i}^{(b)}(t)$ is an upper bound of the level of reflections. $\tilde{\mathrm{h}}_R(t)$ is the MMSE approximation of $|\mathrm{h}_R(t)|$.

$$\tilde{\mathrm{h}}_R(t) \triangleq \mathbb{E}\big(|\mathrm{h}_R(t)|\big) = \begin{cases} 0,\ t < \tau_{qi} \\ 10^{-3\frac{t}{T_{60}}},\ t \geq \tau_{qi}. \end{cases} \tag{3.14}$$

Moreover, since the envelope of a distinct onset is a rising ramp, the delayed reflections are comparatively small. Thus for the duration of the distinct onset,

$$x_i^{(b)}(t) \approx x_{d_i}^{(b)}(t), \tag{3.15}$$

which aligns with the "precedence effect" [132] that speech onsets in microphone signals are dominated by direct-path components.

Therefore from (3.13) and (3.15),

$$\tilde{x}_{R_i}^{(b)}(t) \approx \tilde{\mathrm{h}}_R(t) * \lfloor x_i^{(b)}(t) \rfloor,\ t \geq t_{qi} + \tau_{qi}. \tag{3.16}$$

Note that the upper bound $\tilde{x}_{R_i}^{(b)}(t)$ is independent on the $\hat{v}_{qi}(t)$ term of $\hat{\mathrm{h}}_{qi}(t)$. Thus it can be used as a consistent threshold for detecting distinct onsets in microphone signals $x_i^{(b)}(t)$.

3.3.4 Recursive Averaging for Reflection Level

In practice, signals are observed at a sampling rate of f_s. Using the fact that (3.14) is a low-pass filtering LTI process, a recursive averaging process to approximate (3.16) is proposed:

$$\bar{x}_i^{(b)}[m] = \lambda \cdot \bar{x}_i^{(b)}[m-1] + (1-\lambda) \cdot \lfloor x_i^{(b)}(m/f_s) \rfloor, \tag{3.17}$$

where λ $(0 < \lambda < 1)$ is a forgetting factor.

From (3.17), the recursive averages after the onset arrival can also be rewritten as:

$$\begin{aligned} \bar{x}_i^{(b)}[m] &= (1-\lambda) \sum_{l=m_{qi}}^{m} \lambda^{m-l} \cdot \lfloor x_i^{(b)}(l/f_s) \rfloor \\ &= \mathrm{h}_A[m] * \lfloor x_i^{(b)}(m/f_s) \rfloor,\ m \geq m_{qi} \triangleq \mathrm{round}(t_{qi} \cdot f_s), \end{aligned} \tag{3.18}$$

where the impulse response is:

$$\mathrm{h}_A[m] = \begin{cases} 0, & m < 0 \\ (1-\lambda) \cdot \lambda^m, & m \geq 0. \end{cases} \tag{3.19}$$

Equating the upper bound (3.16) and the recursive averages (3.18),

$$\bar{x}_i^{(b)}[m] = \tilde{x}_{R_i}^{(b)}(m/f_s), \tag{3.20}$$

leads to

$$\mathrm{h}_A[m] \approx \tilde{\mathrm{h}}_R(m/f_s). \tag{3.21}$$

From (3.14) and (3.19),

$$\begin{cases} (1-\lambda)\cdot\lambda^m \approx 0,\ 0 \le m < m_{\tau_{qi}} \\ (1-\lambda)\cdot\lambda^m \approx 10^{-3\frac{m}{T_{60}\cdot f_s}},\ m \ge m_{\tau_{qi}}, \end{cases} \tag{3.22}$$

where $m_{\tau_{qi}} \triangleq \mathrm{round}(\tau_{qi}\cdot f_s)$.

Thus for τ_{qi} larger than a few milliseconds (as discussed in Section 2.3), from (3.22) for $m \ge m_{\tau_{qi}}$, a non-trivial solution for (3.22) is

$$\begin{aligned} \lambda \approx & (1-\lambda)^{-\frac{1}{m}}\cdot 10^{-\frac{3}{T_{60}\cdot f_s}} \\ \approx & 10^{-\frac{3}{T_{60}\cdot f_s}}. \end{aligned} \tag{3.23}$$

When $f_s = 48000$Hz, $\lambda = 0.9998$ for $T_{60} = 0.72$s, and $\lambda = 0.99$ for $T_{60} = 15$ms. It is obvious that λ increases as T_{60} increases (stronger reverberation). As discussed in Section 2.3, T_{60} can be obtained via measurement or estimation [67, 68]. Fig. 3.1 gives an illustration of the recursive averaging for a subband signal. As shown next, the aim is to detect the speech onsets and discard the speech offsets.

3.3.5 Onset Detection

From (3.15) and Appendix B , at the distinct onset

$$x_i^{(b)}(t) \approx \tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t)),\ t \ge t_{qi}, \tag{3.24}$$

where $\tilde{S}_{qi}^{(b)}(t)$ and $\tilde{\phi}_{qi}^{(b)}(t)$ are the envelope and phase defined in (7.19) and (7.20), respectively in Appendix D . Thus from (7.27), an upper bound for

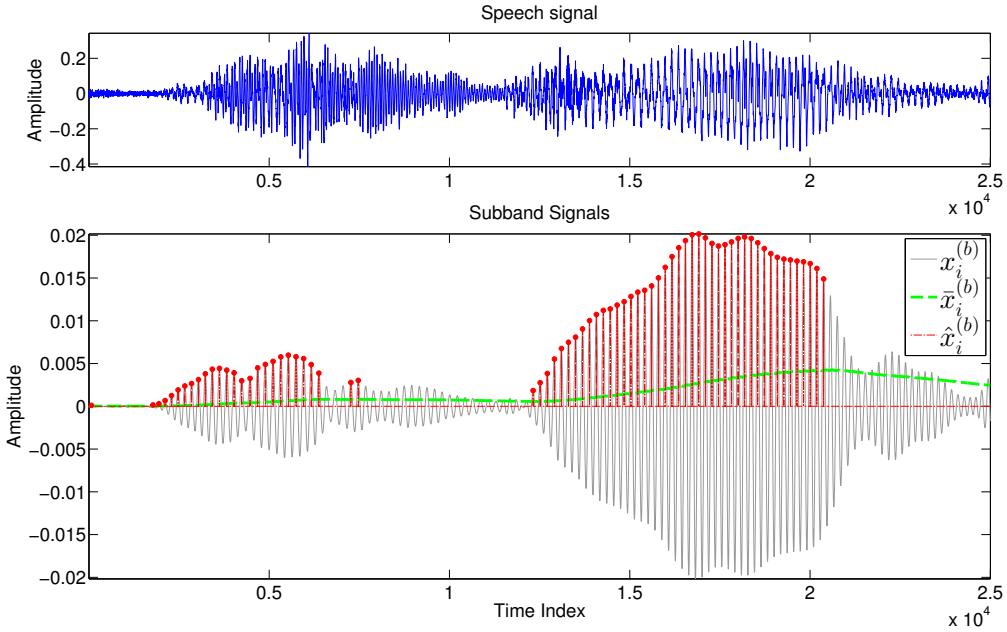

Figure 3.1: Speech signal (top panel), subband signal, recursive average, and encoded subband onset signal (bottom panel). ©2018 IEEE.

its recursive averages is derived, i.e.

$$\frac{\tilde{S}_{qi}^{(b)}(m/f_s)}{\bar{x}_i^{(b)}[m]} \geq \pi. \tag{3.25}$$

Note that the equality in (3.25) holds also when the envelope $\tilde{S}_{qi}^{(b)}(m/f_s)$ is constant. Other parts of signals can be speech offsets or weak utterances corrupted by the reflections. Thus those parts of signals are discarded when (3.25) does not hold.

The envelope limits the peaks of subband signal, i.e.

$$\tilde{S}_{qi}^{(b)}(\hat{m}_{i,n}^{(b)}/f_s) = \lfloor x_i^{(b)}(\hat{m}_{i,n}^{(b)}/f_s)\rfloor, \tag{3.26}$$

where $\hat{m}_{i,n}^{(b)}$ $(n = 1, 2, ...)$ are indices of local peaks

$$\hat{m}_{i,n}^{(b)} = \arg\max_m \lfloor x_i^{(b)}(m/f_s)\rfloor, \ \forall\ m \in (\check{m}_{i,n}^{(b)}, \vec{m}_{i,n}^{(b)}), \tag{3.27}$$

and index pairs $\overleftarrow{m}_{i,n}^{(b)}$ and $\vec{m}_{i,n}^{(b)}$ are consecutive zero-crossings that satisfy

$$\lfloor x_i^{(b)}(m/f_s) \rfloor > 0, \ \forall \ m \in (\overleftarrow{m}_{i,n}^{(b)}, \vec{m}_{i,n}^{(b)}). \tag{3.28}$$

Thus by comparing the local peaks with the recursive averages according to (3.25), we can find the set of indices of distinct onset signals:

$$K_{i+}^{(b)} \triangleq \{m \mid \overleftarrow{m}_{i,n}^{(b)} < m < \vec{m}_{i,n}^{(b)}, \ \frac{\lfloor x_i^{(b)}(\hat{m}_{i,n}^{(b)}/f_s) \rfloor}{\bar{x}_i^{(b)}[\hat{m}_{i,n}^{(b)}]} \geq \pi\}. \tag{3.29}$$

According to (3.15), (3.20) and (3.25), signals with indices $m \in K_{i+}^{(b)}$ are distinct onset signals where direct-path components are dominant, while the rest signals of $\lfloor x_i^{(b)}(m/f_s) \rfloor$ can be corrupted by reflection components and hence are discarded.

3.3.6 Onset Encoding

Once the distinct onsets are found from the subband signals of microphones, they can be encoded to find the locations of speaker q by estimating the TDOA of direct-path sounds (i.e. differences of t_{qi} in (3.8)) between multiple microphones.

Assuming a slow-changing $\phi_q^{(\hbar)}(t)$ in (3.2), the detected distinct onset signals can be rewritten as a convolution:

$$\lfloor x_i^{(b)}(m/f_s) \rfloor \approx \zeta_{\text{cosine}}^{(\hbar,q)}[m] * \sum_{\hat{m}_{i,n}^{(b)} \in \hat{K}_{i+}^{(b)}} x_i^{(b)}(m/f_s) \cdot \delta[m - \hat{m}_{i,n}^{(b)}], \tag{3.30}$$

where $\zeta_{\text{cosine}}^{(\hbar,q)}[m]$ is the non-negative part of the cosine term with peak at $m = 0$,

$$\zeta_{\text{cosine}}^{(\hbar,q)}[m] \triangleq \cos\left(2\pi\hbar f_q m/f_s\right), \ m \in (-\frac{f_s}{4\hbar f_q}, \frac{f_s}{4\hbar f_q}), \tag{3.31}$$

the delta function $\delta[m]$ is defined as

$$\delta[m] = \begin{cases} 1, & m = 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3.32}$$

and $\hat{K}_{i+}^{(b)} \subset K_{i+}^{(b)}$ is the set of indices of onset peaks in (3.27):

$$\hat{K}_{i+}^{(b)} \triangleq \{\hat{m}_{i,n}^{(b)} \mid \frac{\lfloor x_i^{(b)}(\hat{m}_{i,n}^{(b)}/f_s)\rfloor}{\bar{x}_i^{(b)}[\hat{m}_{i,n}^{(b)}]} \geq \pi\}. \tag{3.33}$$

Since the precise timing information of onsets (the signal-scaled delta functions in (3.30)) is crucial to time delay estimation, the slow-changing $\zeta_{\text{cosine}}^{(\hbar,q)}[m]$ term in (3.30) can impair the resolution in location estimates (see e.g. the analysis for PHAT prefiltering in [35]). On the other hand however, sharp peaks can be sensitive to noise and finite observation time [35]. Thus the choice of encoding the $\zeta_{\text{cosine}}^{(\hbar,q)}[m]$ term for cross-correlation represents a compromise between good resolution, accuracy and reliability.

Considering that localization of multiple concurrent speakers requires good resolution, and assuming that the noise is not too strong, the simple encoding can be used by eliminating the $\zeta_{\text{cosine}}^{(\hbar,q)}[m]$ term (or equivalently use $\zeta_{\text{cosine}}^{(\hbar,q)}[m] = \delta[m]$) and encoding the onsets directly with the scaled delta functions in (3.30). The resulting signal (cf. Fig. 3.1) is denoted as $\hat{x}_i^{(b)}[m]$:

$$\hat{x}_i^{(b)}[m] = \sum_{\hat{m}_{i,n}^{(b)} \in \hat{K}_{i+}^{(b)}} x_i^{(b)}(m/f_s) \cdot \delta[m - \hat{m}_{i,n}^{(b)}], \ \forall m \in \mathbb{Z}. \tag{3.34}$$

Some other ways of encoding the signals can be found in the literature that generate spikes in-phase with local signal peaks before cross-correlation, however they were inferred from psychoacoustic observations that the neural spikes are generated by the hair cells in the organ of Corti [39, 133].

3.4 Onset-GSRP and Onset-MCC

In Section 3.3, the distinct onset cues that are dominated by direct-paths have been found. The Onset-MCCC [45] scans through a spatial grid to calculate corresponding MCCCs, and is hence computationally intensive. Based on the Onset-MCCC method, this section proposes the Onset-GSRP localization method using the encoded onset cues.

Assuming that all speech sources are in the far-field, from (2.80), the TDOA (in discrete samples) from a source at $\vec{\wp}(\theta)$ to locations of two microphones $\vec{m}_i$, $\vec{m}_j$ is

$$\Delta^{(ij)}(\vec{\wp}(\theta)) = \text{round}(\tau_{ij}(\theta) \cdot f_s), \tag{3.35}$$

where the location on the azimuthal plane is (using r_s to denote the distance from the candidate source to the array center)

$$\vec{\wp}(\theta) = r_s \cdot [\cos\theta, \sin\theta]. \tag{3.36}$$

For each microphone pair, it is easy to find the range of the relative sample delays over the entire desired space, i.e. $[\Delta^{(ij)}_{min}, \Delta^{(ij)}_{max}]$. For compact array geometries, $|\Delta^{(ij)}_{min}|$ and $|\Delta^{(ij)}_{max}|$ are small. For example, an array of maximum dimension of $d_a = 0.1m$ has maximum possible relative sample delay of about 14, at a sampling frequency of $f_s = 48000$Hz. This saves considerable computational cost. Thus inspired by the inverse mapping idea in [125], this thesis proposes the Onset-GSRP method by calculating the cross-correlations per discrete sample delays within the range $[\Delta^{(ij)}_{min}, \Delta^{(ij)}_{max}]$. Details are explained as follows.

Denote the relationship between the location and the corresponding sample delay (3.35) as a function

$$\Delta^{(ij)}(\vec{\wp}) = \mathcal{M}(\vec{\wp}), \tag{3.37}$$

where $\mathcal{M}(\cdot)$ stands for a mapping function, with an inverse function $\mathcal{M}^{-1}(\cdot)$. For a single microphone pair, there is ambiguity in the mapping that one particular sample delay may correspond to two azimuthal locations. However,

this is resolved by the use of multiple microphone pairs. Thus the set of locations that correspond to these discrete sample delays can be denoted as

$$\hat{\wp}^{(ij)} = \{\overrightarrow{\wp} \mid \overrightarrow{\wp} = \mathcal{M}^{-1}(\Delta^{(ij)}),\ \forall\ \Delta^{(ij)} \in [\Delta^{(ij)}_{min}, \Delta^{(ij)}_{max}]\}. \tag{3.38}$$

Meanwhile, corresponding to these sample delays (and hence the locations), the cross-correlation coefficients can also be found for a frame length of N, i.e.

$$\begin{aligned} e^{(b)}_{ij}[k, \overrightarrow{\wp}] &= \text{xcorr}\big(\hat{x}^{(b)}_i[m], \hat{x}^{(b)}_j[m - \Delta^{(ij)}(\overrightarrow{\wp})]\big) \\ &= \frac{\sum_{m=kM-N+1+\Delta^{(ij)}}^{kM} \tilde{x}^{(b)}_i[m] \cdot \tilde{x}^{(b)}_i[m - \Delta^{(ij)}]}{\sqrt{\sum_{m=kM-N+1}^{kM} [\tilde{x}^{(b)}_i[m]]^2 \cdot \sum_{m=kM-N+1+\Delta^{(ij)}}^{kM} [\tilde{x}^{(b)}_i[m - \Delta^{(ij)}]]^2}},\ \forall\ \overrightarrow{\wp} \in \hat{\wp}^{(ij)}, \end{aligned} \tag{3.39}$$

where $\tilde{x}^{(b)}_i[\cdot]$ is $\hat{x}^{(b)}_i[\cdot]$ with DC offset removed. Note that by definition the cross-correlation coefficients $|e^{(b)}_{ij}[k, \overrightarrow{\wp}]| \in [0, 1]$.

These cross-correlation coefficients are then linearly interpolated across the spatial grid. For example, if only the azimuthal source DOAs across $[0^\circ, 360^\circ)$ are desired and estimated at 1° grid steps, the locations can be simply denoted as $\overrightarrow{\wp}_m$, indexed by the DOA. Consequently, between any two consecutive locations $\overrightarrow{\wp}_{m_a}$ and $\overrightarrow{\wp}_{m_b}$, $(m_a \neq m_b)$ with non-zero cross-correlation coefficients,

$$e^{(b)}_{ij}[k, \overrightarrow{\wp}_m] = e^{(b)}_{ij}[k, \overrightarrow{\wp}_{m_a}] \frac{m - m_b}{m_a - m_b} + e^{(b)}_{ij}[k, \overrightarrow{\wp}_{m_b}] \frac{m_a - m}{m_a - m_b},\ m \in (m_b, m_a). \tag{3.40}$$

The same process repeats for all the microphone pairs, excluding those far-apart to avoid spatial alias. Then for each point of location $\overrightarrow{\wp}_m$, the cross-correlation coefficients are accumulated, i.e.

$$e^{(b)}[k, \overrightarrow{\wp}_m] = \sum_{(i,j) \in \mathbf{P}^{(b)}} e^{(b)}_{ij}[k, \overrightarrow{\wp}_m], \tag{3.41}$$

where microphone pairs are selected to avoid spatial alias [39], which fol-

lows (3.42), where only $i < j$ pairs are used without duplication, i.e.

$$\mathbf{P}^{(b)} = \{(i,j) \mid \|\vec{m}_i - \vec{m}_j\| < 2\pi v/[f_s(\omega_c^{(b)} + 2\omega_B^{(b)})],\ i < j\}. \tag{3.42}$$

Finally all the subband results are accumulated to obtain the localization function, i.e.

$$\epsilon^{\text{onset-gsrp}}[k, \vec{\wp}_m] = \frac{1}{N_b}\sum_{b=1}^{N_b} e^{(b)}[k, \vec{\wp}_m]. \tag{3.43}$$

An intuitive alternative to the Onset-GSRP is by simply replacing the summation in (3.41) with the product operator, and the resulting localization function still produces peaks corresponding to speaker locations, i.e.

$$\epsilon^{\text{onset-mcc}}[k, \vec{\wp}_m] = \frac{1}{N_b}\sum_{b=1}^{N_b}\prod_{(i,j)\in\mathbf{P}^{(b)}} e_{ij}^{(b)}[k, \vec{\wp}_m]. \tag{3.44}$$

This new algorithm is referred to as the Onset-MCC (to make a distinction from the Onset-MCCC method in [45]) which, as will be explained in Section 3.5 and evaluated in Section 3.7, produces reverberation-robust localization results with the improved DOA resolution.

For compact microphone arrays with a maximal dimension d_a, the TDOAs between microphones are considered independent of r_s when $r_s \gg d_a$. Thus a fixed value of r_s can be assumed to scan the DOA θ [39, 45].

This chapter considers only azimuthal DOAs of speakers. The proposed method however, can be easily extended for estimating Cartesian locations of speakers in the azimuthal plane and 3D space using multiple microphone arrays.

3.5 Redundant Information and MCC-PHAT

The RIR model in (3.3) may be restrictive by assuming that the source is not located too close to reflecting surfaces, although it has enabled the derivation and led to a working solution. Here a more general RTF model is used and a new reverberation-robust localization approach is proposed by exploiting the

redundant information from the microphone array. Note that for notational simplicity, time frame indices of STFT signals are not written explicitly.

3.5.1 RTF Model

The RTF as described in (2.35) of Section 2.3 can be written as

$$\mathrm{H}_{qi}(\theta_q, \omega) = \mathrm{d}_{qi}(\omega) + \hat{v}_{qi}(\omega) + \tilde{v}_{qi}(\omega), \tag{3.45}$$

where

$$\text{the direct-path } \mathrm{d}_{qi}(\omega) \triangleq \mathrm{e}^{-\jmath\omega f_s t_{d_{qi}}(\theta_q)} \tag{3.46a}$$

$$\text{early-reflections } \hat{v}_{qi}(\omega) \triangleq \sum_{\bar{q}} \hat{v}_{\bar{q}i} \mathrm{e}^{-\jmath\omega f_s t_{d_{qi}}(\theta_{\bar{q}i})}, \tag{3.46b}$$

$\theta_{\bar{q}i}$ denotes the set of DOAs from early reflections, $\hat{v}_{\bar{q}i} \in \mathbb{R}$ ($|\hat{v}_{\bar{q}i}| < 1$) the magnitude, and $\tilde{v}_{qi}(\omega) \in \mathbb{C}$ the RTF of the diffuse reflections. Early reflections are considered directional (provided with perfectly smooth reflection surfaces, cf. (2.17), otherwise only part of reflections are directional), and in general stronger than the ensuing diffuse reflections (cf. Fig. 2.3). It is also reasonable to assume that diffuse reflection responses are zero-mean and uncorrelated [61] and spatially white, i.e.

$$\mathbb{E}[\tilde{v}_{qi}(\omega)] = 0 \tag{3.47a}$$

$$\mathbb{E}[\tilde{v}_{qi}(\omega)\tilde{v}^{\star}_{qj}(\omega)] = \sigma^2_{\tilde{v}}(\omega)\delta_i(j), \tag{3.47b}$$

where $\delta_i(j)$ is a Kronecker delta function, i.e.

$$\delta_i(j) = \begin{cases} 0, \text{ if } i = j \\ 1, \text{ else,} \end{cases} \tag{3.48}$$

and according to the above assumptions,

$$0 < \sigma^2_{\tilde{v}}(\omega) \ll \sigma^2_{\hat{v}_{\bar{q}}}(\omega) < 1, \tag{3.49}$$

where $\sigma^2_{\hat{v}_{\bar{q}}}(\omega) \triangleq \hat{v}_{\bar{q}i}(\omega) \cdot \hat{v}_{\bar{q}j}(\omega)$, and for a compact microphone array, $\hat{v}_{\bar{q}i}(\omega) \approx \hat{v}_{\bar{q}j}(\omega)$.

Also assume that noise signals are zero-mean, uncorrelated and white, and much weaker than the speech signal, i.e.

$$\mathbb{E}[N_i(\omega)] = 0 \tag{3.50a}$$

$$\mathbb{E}[N_i(\omega)N_j^\star(\omega)] = \sigma_N^2(\omega)\delta_i(j), \tag{3.50b}$$

where the noise variance $\sigma_N^2(\omega) \ll \sigma_S^2(\omega)$, the speech signal power is denoted $\sigma_S^2(\omega)$, and $\hat{\sigma}_S^2(\omega)$ denotes a short-time estimate of the speech signal power. Unless otherwise noted hereafter, the mathematical expectations of STFT domain variables are approximated with short-time averages over time frames, under the WSS assumption.

3.5.2 Direct-path to Reflection Ratio

The classical GCC-PHAT method [35] has been described in Section 2.5.3. From (2.81), (2.82) and (2.83), the GCC-PHAT estimator is rewritten here

$$\epsilon_{ij}^{\text{gcc-phat}}(\tau_{ij}(\theta)) = \sum_{\omega} \Psi_{ij}(\omega) \cdot G_{x_i x_j}(\omega) \cdot e^{\jmath\omega f_s \tau_{ij}(\theta)}, \tag{3.51}$$

where $\Psi_{ij}(\omega) = |G_{x_i x_j}(\omega)|^{-1}$, and when the sound of speaker q is dominant,

$$\begin{aligned}
&G_{x_i x_j}(\omega) = \mathbb{E}[X_i(\omega)X_j^\star(\omega)] \\
=&\hat{\sigma}_S^2(\omega)\{e^{-\jmath\omega f_s \tau_{ij}(\theta_q)} + \sum_{\bar{q}} \sigma^2_{\hat{v}_{\bar{q}}}(\omega) e^{-\jmath\omega f_s [t_{d_{qi}}(\theta_{\bar{q}i}) - t_{d_{qj}}(\theta_{\bar{q}j})]}\} \\
&+ \hat{\sigma}_S^2(\omega) \sum_{\bar{q}} \sigma_{\hat{v}_{\bar{q}}}(\omega) e^{-\jmath\omega f_s [t_{d_{qi}}(\theta_q) - t_{d_{qj}}(\theta_{\bar{q}j})]} \\
&+ \mathbb{E}[O_{\text{cross-terms}}(\tilde{v}_{qi}(\omega), \tilde{v}_{qj}^\star(\omega), N_i(\omega), N_j^\star(\omega))] \\
\approx&\hat{\sigma}_S^2(\omega)\{e^{-\jmath\omega f_s \tau_{ij}(\theta_q)} + \sum_{\bar{q}} \sigma_{\hat{v}_{\bar{q}}}(\omega) e^{-\jmath\omega f_s [t_{d_{qi}}(\theta_q) - t_{d_{qj}}(\theta_{\bar{q}j})]} + \sum_{\bar{q}} \sigma^2_{\hat{v}_{\bar{q}}}(\omega) e^{-\jmath\omega f_s \tau_{ij}(\theta_{\bar{q}})}\},
\end{aligned} \tag{3.52}$$

where from (3.47) and (3.50), the cross-terms of noise and diffuse reflections in (3.52) are negligible. For a closely located microphone pair, $\theta_{\bar{q}i} = \theta_{\bar{q}j} \triangleq \theta_{\bar{q}}$,

hence in the last line of (3.52), $\tau_{ij}(\theta_{\bar{q}}) = t_{d_{qi}}(\theta_{\bar{q}i}) - t_{d_{qj}}(\theta_{\bar{q}j})$.

Further assume that the early reflections do not considerably vary the spectral distribution of the speech signal. Then the terms $\Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)$ and $\Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)\sigma_{\hat{v}_{\bar{q}}}^2(\omega)$ can be treated as constants (i.e. pre-whitening weights) for each frequency when the noise is not too strong in each time frame. Therefore, using (3.45), (3.51) and (3.52), it can be easily seen that

$$
\begin{aligned}
&\epsilon_{ij}^{\text{gcc-phat}}(\tau_{ij}(\theta)) \\
&\approx \underbrace{\sum_{\omega} \Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega) e^{\jmath\omega f_s \tau_{ij}(\theta)} e^{-\jmath\omega f_s \tau_{ij}(\theta_q)}}_{\text{direct-path}} \\
&\quad + \underbrace{\sum_{\omega,\bar{q}} \Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)\sigma_{\hat{v}_{\bar{q}}}^2(\omega) e^{\jmath\omega f_s \tau_{ij}(\theta)} e^{-\jmath\omega f_s \tau_{ij}(\theta_{\bar{q}})}}_{\text{early reflections}} \\
&\quad + \underbrace{\sum_{\omega,\bar{q}} \Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)\sigma_{\hat{v}_{\bar{q}}}(\omega) e^{\jmath\omega f_s \tau_{ij}(\theta)} e^{-\jmath\omega f_s [t_{d_{qi}}(\theta_q) - t_{d_{qj}}(\theta_{\bar{q}j})]}}_{\text{cross-terms of direct-path and early reflections}} \\
&= \underbrace{\delta_\theta(\theta_q) \sum_{\omega} \Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)}_{\text{direct-path}} + \underbrace{\sum_{\bar{q}} \delta_\theta(\theta_{\bar{q}}) \sum_{\omega} \Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)\sigma_{\hat{v}_{\bar{q}}}^2(\omega)}_{\text{early reflections}}.
\end{aligned}
\tag{3.53}
$$

where the cross-terms of the direct-path and early reflections are negligible, since there is no non-trivial DOA solution to $\tau_{ij}(\theta) - [t_{d_{qi}}(\theta_q) - t_{d_{qj}}(\theta_{\bar{q}j})] = 0$ for a closely located microphone pair. Thus the localization function of the GCC-PHAT produces a peak at θ_q, and also spurious peaks at $\theta_{\bar{q}}$ due to early reflections, besides the negligible contributions by cross-terms.

Define the direct-path to reflection ratio (DRR) as the ratio of the dominant direct-path peak to the strongest reflection peak. Thus from (3.53),

$$
\text{DRR}^{\text{gcc-phat}} = \frac{1}{\max\limits_{\bar{q}}(\sigma_{\hat{v}_{\bar{q}}}^2)}. \tag{3.54}
$$

To improve the DRR, the well-accepted SRP-PHAT is one of the possible approaches. From (2.85) and (3.53), the SRP-PHAT concentrates the direct-

path peaks at θ_q, i.e.

$$\begin{aligned}
&\epsilon^{\text{srp-phat}}(\theta) = \sum_{i}^{I_M}\sum_{j}^{I_M}\epsilon_{ij}^{\text{gcc-phat}}(\tau_{ij}(\theta)) \\
&\approx \underbrace{\sum_{i}^{I_M}\sum_{j}^{I_M}\sum_{\omega}\Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)e^{\jmath\omega f_s\tau_{ij}(\theta)}e^{-\jmath\omega f_s\tau_{ij}(\theta_q)}}_{\text{direct-path}} \\
&\quad + \underbrace{\sum_{i}^{I_M}\sum_{j}^{I_M}\sum_{\omega,\bar{q}}\Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)\sigma_{\hat{v}_{\bar{q}}}^2(\omega)e^{\jmath\omega f_s\tau_{ij}(\theta)}e^{-\jmath\omega f_s\tau_{ij}(\theta_{\bar{q}})}}_{\text{early reflections}} \\
&= \underbrace{\delta_\theta(\theta_q)\sum_{i}^{I_M}\sum_{j}^{I_M}\sum_{\omega}\Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)}_{\text{direct-path}} + \underbrace{\sum_{i}^{I_M}\sum_{j}^{I_M}\sum_{\bar{q}}\delta_\theta(\theta_{\bar{q}})\sum_{\omega}\Psi_{ij}(\omega)\hat{\sigma}_S^2(\omega)\sigma_{\hat{v}}^2(\omega)}_{\text{early reflections}}.
\end{aligned} \tag{3.55}$$

Note in (3.55) that there are I_M^2 microphone pairs used from a total of I_M microphones. This apparently indicates redundant information. Why would SRP-PHAT work better than GCC-PHAT in reverberant environments? How could the redundant information be useful?

Based on the assumptions that all the microphones are spatially close and the early reflections are perfectly directional, the localization function of each microphone pair produces spurious peaks due to directional early reflections, which perfectly align. Thus it is basically the same case with the GCC-PHAT in that the DRR does not change, i.e. $\text{DRR}^{\text{srp-phat}} = \text{DRR}^{\text{gcc-phat}}$. In practice however, the reflection surfaces may not be perfectly smooth, and the reflection coefficients may not be incident-angle-independent or frequency-independent, hence not all of the spurious peaks from the localization functions of different microphone pairs (induced by the reflection) may align at a certain DOA. Thus the resulting DRR from (3.55) is

$$\text{DRR}^{\text{srp-phat}} \geq \text{DRR}^{\text{gcc-phat}} = \frac{1}{\max\limits_{\bar{q}}(\sigma_{\hat{v}_{\bar{q}}}^2)}, \tag{3.56}$$

which implies that the SRP-PHAT localization method may improve the DRR

using the redundant information from all microphones.

3.5.3 Redundant Information

By exploiting the redundant information, another approach to improve DRR is via multiplication of the GCC-PHAT functions, instead of the summation as the SRP-PHAT does in (2.85). For example, using three microphones i, j, k,

$$\begin{aligned}
&\epsilon^{\text{multip}}(\theta) = \prod_{\{ij\}\in\{ij,jk,ik\}} \epsilon_{ij}^{\text{gcc}-\text{phat}}(\tau_{ij}(\theta)) \\
&\approx \underbrace{\sum_{\omega}\prod_{ij}\Psi_{ij}(\omega)\hat{\sigma}_S^6(\omega)e^{J\omega f_s\left[\tau_{ij}(\theta)+\tau_{jk}(\theta)+\tau_{ik}(\theta)-\tau_{ij}(\theta_q)-\tau_{jk}(\theta_q)-\tau_{ik}(\theta_q)\right]}}_{\text{direct-path}} \\
&\quad + \underbrace{\sum_{\bar{q}}\sum_{\omega}\prod_{ij}\Psi_{ij}(\omega)\hat{\sigma}_S^6(\omega)\sigma_{\hat{v}_{\bar{q}}}^6(\omega)e^{J\omega f_s[\tau_{ij}(\theta)+\tau_{jk}(\theta)+\tau_{ik}(\theta)-\tau_{ij}(\theta_{\bar{q}})-\tau_{jk}(\theta_{\bar{q}})-\tau_{ik}(\theta_{\bar{q}})]}}_{\text{early reflections}} \\
&= \underbrace{\delta_\theta(\theta_q)\sum_{\omega}\prod_{ij}\Psi_{ij}(\omega)\hat{\sigma}_S^6(\omega)}_{\text{direct-path}} + \underbrace{\sum_{\bar{q}}\delta_\theta(\theta_{\bar{q}})\sum_{\omega}\prod_{ij}\Psi_{ij}(\omega)\hat{\sigma}_S^6(\omega)\sigma_{\hat{v}_{\bar{q}}}^6(\omega)}_{\text{early reflections}}.
\end{aligned} \tag{3.57}$$

Therefore, from (3.57), the DRR becomes $\text{DRR}^{\text{multip}} = [\frac{1}{\max_{\bar{q}}(\sigma_{\hat{v}_{\bar{q}}}^2)}]^3$, and it is straightforward to see that for I_M microphones,

$$\text{DRR}^{\text{multip}} = [\frac{1}{\max_{\bar{q}}(\sigma_{\hat{v}_{\bar{q}}}^2)}]^{C_{I_M}^2}, \tag{3.58}$$

where the combination $C_{I_M}^2 \triangleq I_M \cdot (I_M - 1)/2$. This is a significant improvement from (3.54), since $\sigma_{\hat{v}_{\bar{q}}}^2 < 1$ from (3.49).

3.5.4 MCC-PHAT

Following the motivation of exploiting the redundant information, a multi-channel extension to the GCC-PHAT is formulated[2]

$$\epsilon^{\text{mcc}-\text{phat}}(k,\theta) \triangleq \prod_{(i,j)\in\mathbf{P}} \epsilon_{ij}^{\text{gcc}-\text{phat}}(k,\tau_{ij}(\theta)), \tag{3.59}$$

where the set of microphone pairs $\mathbf{P}$ includes all microphone pairs with $i<j$. For implementation, the real part of $\epsilon_{ij}^{\text{gcc}-\text{phat}}$ is used, and microphone pairs that are far apart are eliminated as in (3.60), which is a trade-off between localization performance and computational efficiency.

$$\mathbf{P} = \{(i,j) \mid \|\vec{m}_i - \vec{m}_j\| < v/f_{max});\ i<j\}, \tag{3.60}$$

where f_{max} is the maximum signal frequency considered.

From (2.83) and similar to (3.44), when θ matches a speaker DOA at time frame k, $\epsilon_{ij}^{\text{gcc}-\text{phat}}$ would be maxima, hence a maximum of $\epsilon^{\text{mcc}-\text{phat}}(k,\theta)$. This extension of the GCC-PHAT method is referred to as the MCC-PHAT method. When there are only two closely placed microphones, the MCC-PHAT becomes the GCC-PHAT. Compared with the SRP-PHAT, the MCC-PHAT provides superior resolution. This is easy to understand, as the SRP-PHAT can be viewed as the MCC-PHAT with logarithmic scale, and each term in the summation is scaled first logarithmically. The logarithmic scaling suppresses better those values that are close to zero (indicating a low probability of source existence), while still concentrating the peaks due to source DOAs, thus providing better resolution. Detailed test results will be shown in Section 3.7.1. Similarly, the summation in (3.41) of the Onset-GSRP can also be replaced with the product operator, which leads to the Onset-MCC method as will also be evaluated in Section 3.7.1.

[2]Although originally inspired by the Onset-MCCC, there are similarities between the proposed MCC-PHAT and the already popular SRP-PHAT. Hence in the performance studies, the two methods are compared, amongst other methods.

3.6 DOA Estimates Extraction

Using the TF sparsity assumption for speech signals, the peaks of localization functions may not appear at the same time, even for concurrent speakers. Considering that the speakers cannot move too fast in an enclosed environment, temporal averaging of length $t_{avg} > 0$ with time shift of $t_{shift} \in (0, t_{avg}]$ is used for the smoothed localization function in each time frame,

$$\bar{\epsilon}(k,\theta) = \frac{1}{f_s \cdot t_{avg}} \sum_{k'=k\cdot f_s\cdot t_{shift}-f_s\cdot t_{avg}+1}^{k\cdot f_s\cdot t_{shift}} \epsilon(k',\theta). \tag{3.61}$$

Peaks of $\bar{\epsilon}(k,\theta)$ correspond to candidate DOA estimates of active speakers. For an unknown number of concurrent speakers, distinct local peaks as in (3.62) and (3.63) are selected. Define $\hat{\Theta}_k$ as the set of DOA estimates at time frame k that correspond to local peaks of $\bar{\epsilon}(k,\cdot)$, cf. (2.62),

$$\hat{\Theta}_k = \{\hat{\theta}_{i_k} \mid i_k = 1,\ldots,N_k\}, \tag{3.62}$$

where $\hat{\theta}_{i_k}$ satisfies $\bar{\epsilon}(k,\hat{\theta}_{i_k}) \geq T_{\bar{\epsilon}}$ and

$$\hat{\theta}_{i_k} = \arg\max_{\theta_{i_k}} \bar{\epsilon}(k,\theta_{i_k}),\ \forall\theta_{i_k} \in [\hat{\theta}_{i_k} - \theta_r,\ \hat{\theta}_{i_k} + \theta_r], \tag{3.63}$$

integer N_k is the number of estimated speakers at time frame k. $T_{\bar{\epsilon}} \in \mathbb{R}$ is an empirical threshold that can be calibrated as the valid range of $\bar{\epsilon}$ depends on the geometry of the microphone array, the microphones used as well as the noise and interferences. Parameter θ_r ($0 < \theta_r \leq 180^\circ$) indicates the minimum angular separation of a DOA estimator. A θ_r that is too small can result in clutters of location estimates, while a θ_r that is too large can cause miss-detections of DOAs. All angles are wrapped into $[0^\circ, 360^\circ)$. When there is no valid peak found from (3.63), $N_k = 0$ and $\hat{\Theta}_k = \emptyset$.

3.7 Numerical Studies

This section compares the performance of the proposed Onset-GSRP, Onset-MCC, and MCC-PHAT, with the Neuro-Fuzzy [39], the SRP-PHAT, the MUSIC, the TF-CHB [34] and the EB-ESPRIT [40] methods, not only in simulated reverberant and noisy conditions, but also in a real office room ($T_{60} \approx 0.65$s). The uniform circular microphone array is studied for its rotational symmetry. A circular microphone array with $I_M = 8$ equidistant omnidirectional elements and radius $r_a = 0.05$m is placed horizontally in the test environment, and speaker DOAs on the same azimuthal plane is evaluated. The microphone signals are sampled synchronously at a sampling rate of 48000Hz. Note that it is straightforward to apply the proposed methods to other compact array geometries.

3.7.1 Experimental Set-up

Choose $r_s = 1$m in (3.36) and scan the DOA in 1° steps. The angular separation θ_r in (3.63) is chosen as 30° unless otherwise noted. The snapshot frame length is 20ms. For temporal averaging in (3.61) the length is $t_{avg} = 0.5$s. Here the TF-CHB, MUSIC and EB-ESPRIT formulate covariance matrices over 20ms time segments. The TF-CHB, MUSIC and EB-ESPRIT do not require high sampling rate, thus signals are resampled at 8000Hz, and accordingly, they use frequencies up to 4000Hz.

For all tests of the Onset-GSRP and Onset-MCC methods here, $\lambda = 0.9998$ in (3.17), which corresponds to $T_{60} \approx 0.72$s as in (3.23). The gammatone filter [46,57,58] as the subband filter in (3.5) is used for its linear phase and frequency selectivity:

$$g^{(b)}(t) = (t + t_d)^{\vartheta - 1} e^{-2\pi f_b^{(b)}(t + t_d)} \cos(2\pi f_c^{(b)} t), \; t \geq -t_d. \tag{3.64}$$

Here integer ϑ is the order of filter ($\vartheta = 4$ here), t_d is time delay for alignment between filter bands, $f_b^{(b)}$ scaling factor for the bandwidth [46,57], and $f_c^{(b)}$ the center frequency of each subband chosen on the equivalent rectangular

bandwidth-rate scale (ERBS) [46]. The center frequencies of the filterbank range from 250Hz to 3600Hz, and the number of subbands is $N_b = 16$. The maximum frequency for MCC-PHAT is $f_{max} = 4000$Hz.

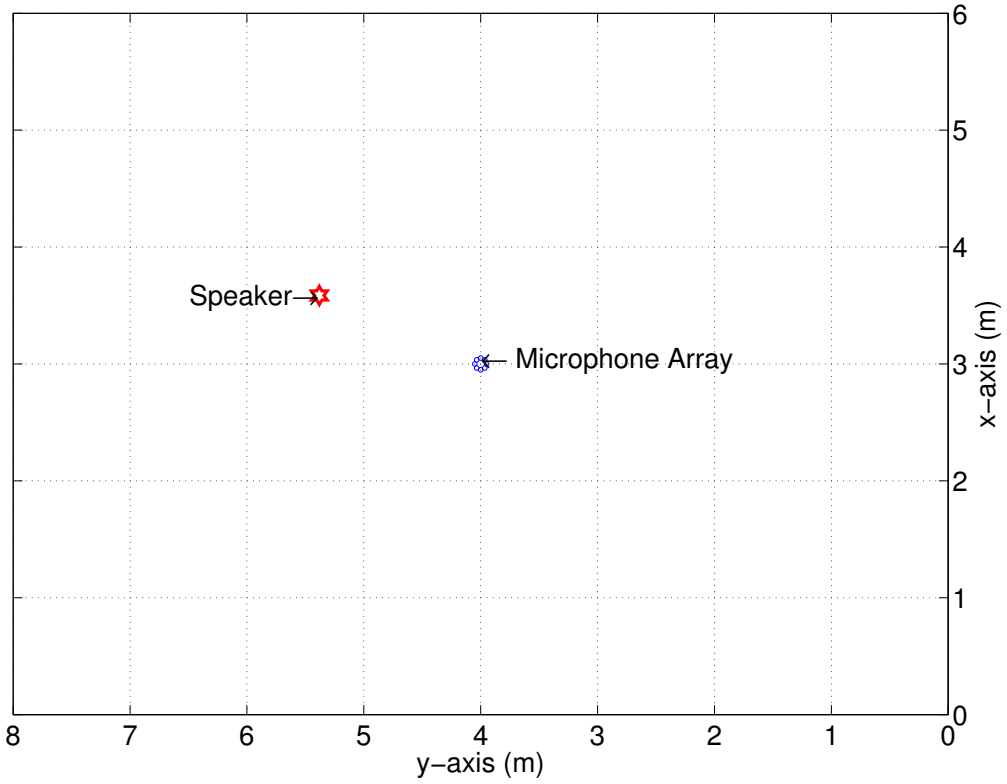

Figure 3.2: A static source not close to wall (DOA is 67°).

The evaluations of five scenarios as follows are presented. Varying reverberation and additive white noise are applied. The reverberation time T_{60} of the simulated environments ranges from 0.2 to 1s simulated using the image-source method (ISM) [62, 140]. Additive uncorrelated Gaussian white noise is applied to each microphone and the signal-to-noise ratio (SNR) varies up to 10dB.

SCENARIO 1 - A STATIC SOURCE NOT TOO CLOSE TO WALL (SIMULATION)

As shown in Fig. 3.2, the direction of the x-axis is defined as the 0° DOA. The UCA is placed at the center of a rectangular room of 6m×8m×3m (width ×

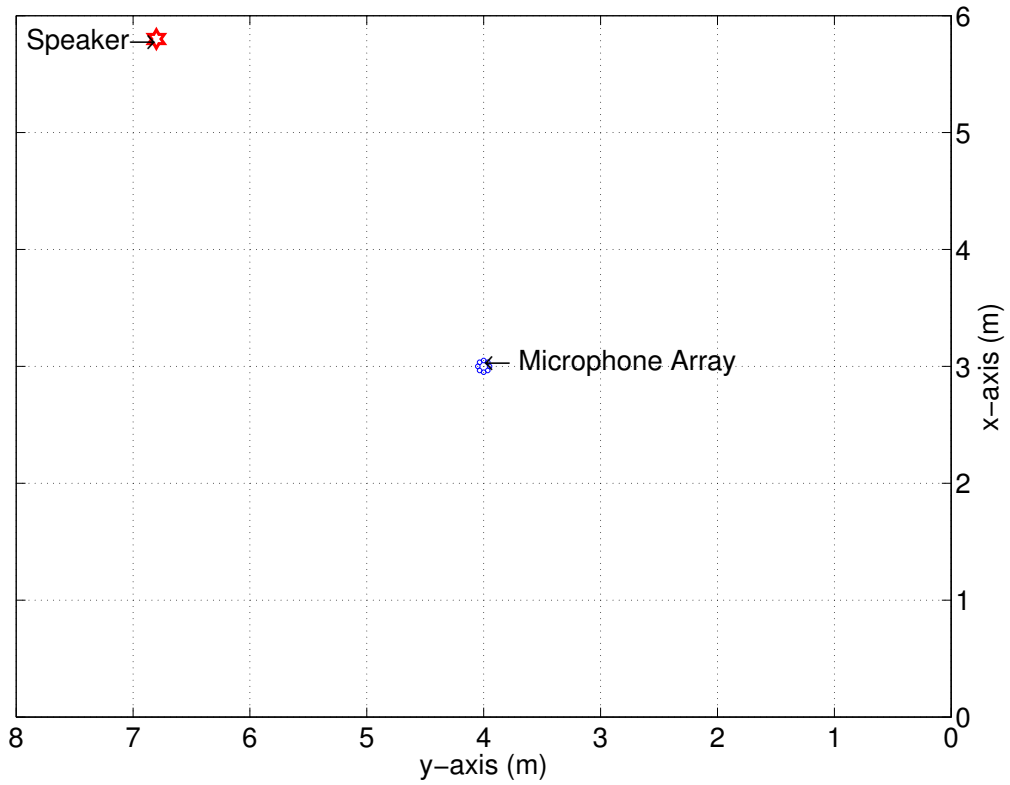

Figure 3.3: A static source close to wall (DOA is 45°).

length × height). The speaker DOA is 45°, and the distance to the closest wall is more than 1m. The speech segment used is 4 seconds long.

SCENARIO 2 - A STATIC SOURCE CLOSE TO WALL (SIMULATION)

Following the discussion in Section 2.3.2, most of the methods can work well for the localization of a single speaker that is not located close to an acoustically reflective object. In this scenario, the more challenging case is evaluated when a single static source is located close to the wall. The room and microphone set-up is the same as in Scenario 1. As shown in Fig. 3.3, the speaker DOA is 45°, and the distance to the closest wall is fixed to 0.2m. The speech segment used is 4 seconds long.

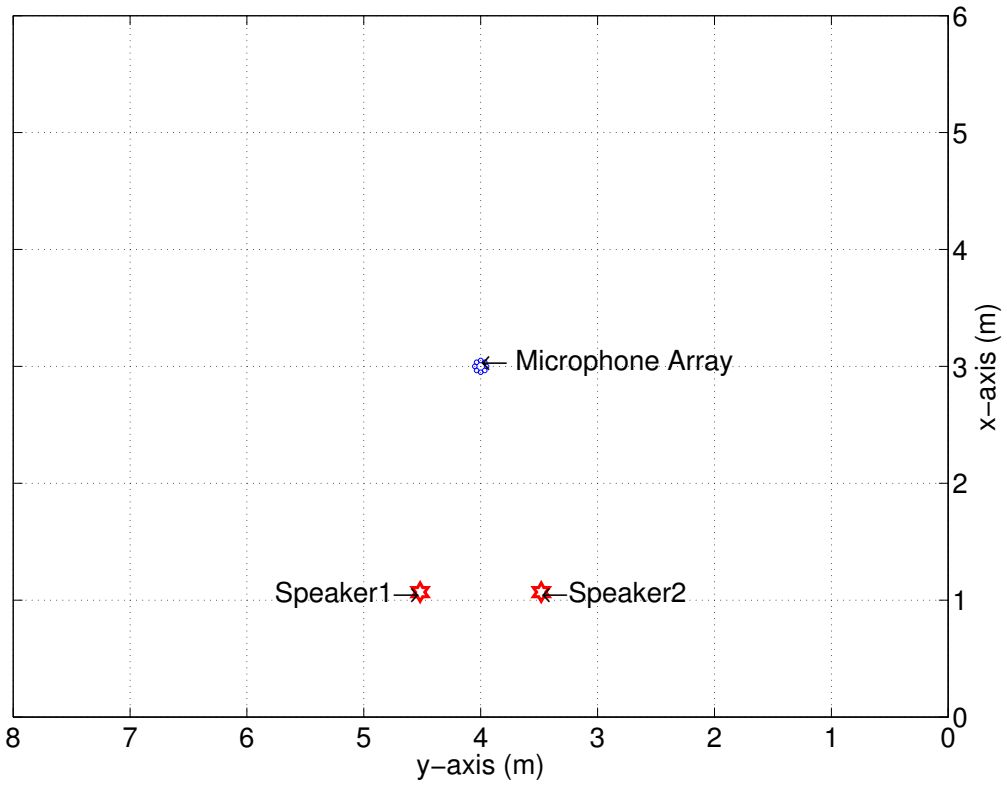

Figure 3.4: Two static sources of close DOAs.

SCENARIO 3 - TWO STATIC SOURCES (SIMULATION)

Fig. 3.4 shows the test set-up for the scenario of two static concurrent speakers. The DOA resolution using respective localization methods is studied. The room and microphone set-up is the same as in Scenario 1. Two speakers are located 2m away from microphone array center. Concurrent speech signals lasts for 4 seconds, each with the same averaged power. Two cases are tested, with DOAs of 170° and 190°, and 165° and 195° respectively. Hence the DOA distances between the two speakers are 20° and 30°, respectively. The angular separation θ_r in (3.63) is chosen as 15° in this scenario.

SCENARIO 4 - THREE STATIC SOURCES (SIMULATION)

As shown in Fig. 3.5, three different static speakers (Speaker1, male; Speaker2, female; Speaker3, male) talk concurrently in this scenario, each with a speech

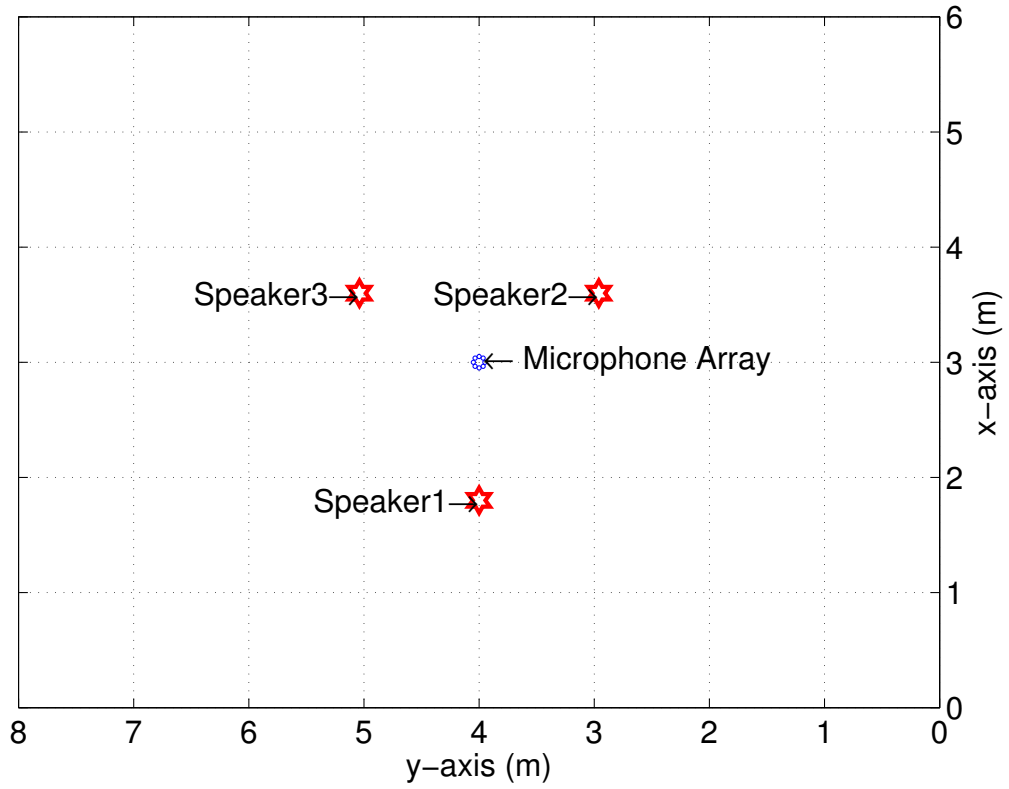

Figure 3.5: Top view of room and set-up (simulation). Locations of microphones and speakers are respectively in circles and stars. ©2018 IEEE.

segment of 4 seconds and the same averaged source power. The room and microphone set-up is the same as in Scenario 1. The speakers locate at DOAs of 180° (Speaker1), 300° (Speaker2) and 60° (Speaker3), respectively, all at a distance of 1.2m from the center of the microphone array.

SCENARIO 5 - THREE MOVING SOURCES (REAL-WORLD)

In this case, three speakers are moving while talking. The experiment is carried out in a real office room with measured reverberation time of $T_{60} \approx 0.65$s. As shown in Fig. 3.6, the dimensions of the room are 3.4m$\times$7.8m$\times$2.7m (width $\times$ length $\times$ height). Equipment used include the RME^{TM} OctaMic XTCs as microphone pre-amplifiers and the HDSPe MADI FX for data acquisition, as well as omnidirectional electrect microphones. Microphones are

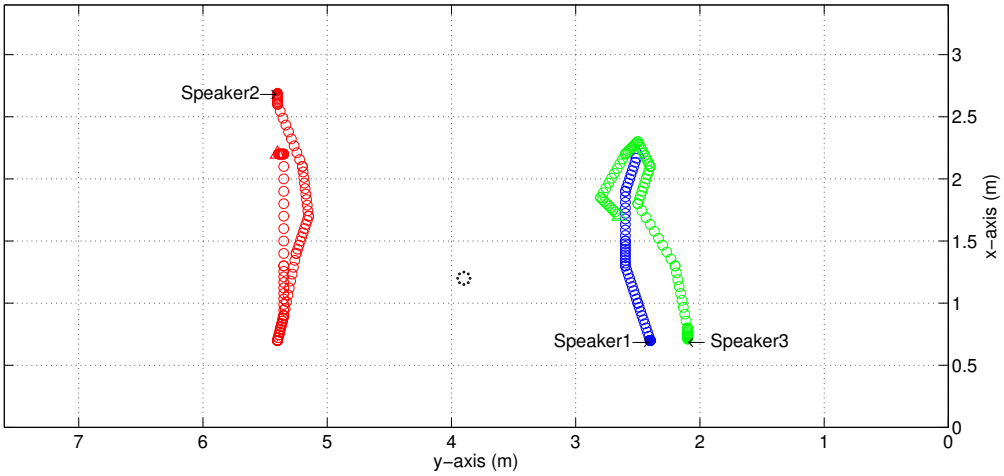

Figure 3.6: Top view of room and set-up (real-world). Locations of microphones are in black circles. Tracks of moving speakers in blue (Speaker1), red (Speaker2) and green (Speaker3). Starting locations of tracks are solid circles and ending locations are triangles. ©2018 IEEE.

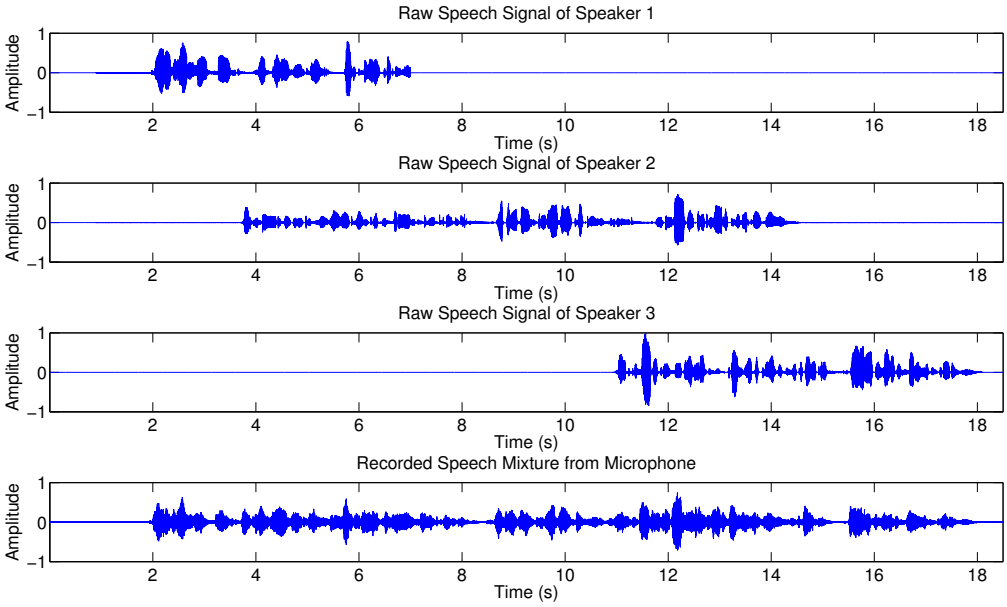

Figure 3.7: Raw signals of moving speakers (top three panels) and a normalized real recording from one of the microphones in the real reverberant room (bottom panel). ©2018 IEEE.

connected to RME^{TM} OctaMic XTC, and HDSPe MADI FX acquires signals from multiple channels of the latter. A camera is used to record the ground

Figure 3.8: Test room set-up.

truth of speakers' movement and sound to compare with estimated results.

The circular microphone array is placed close to the center of the room

at $[1.2, 3.9, 1.5]$m. Omnidirectional electret microphones are used. Three moving speakers talk and move in a random sequence. Speaker signals are chosen from the TIMIT database [141]. The trajectories of speakers are also plotted in Fig. 3.6, with different colors. Fig. 3.7 shows the waveforms of speech signals and their starting and ending time, as well as the real recording from one of the microphones in the reverberant room. As shown in Fig. 3.8, the author of the thesis, has also spent the time and effort to procure equipment, clean up and set up the room for recordings and experiment.

3.7.2 Test Results

SCENARIO 1 - A STATIC SOURCE NOT CLOSE TO WALL (SIMULATION)

Fig. 3.9 and Fig. 3.10 provide the normalized (and scaled by $10\lg(\cdot)$) DOA estimation histograms of localization functions from the proposed Onset-GSRP, Onset-MCC and MCC-PHAT methods as well as that of the Neuro-Fuzzy method, the steered-response power of the TF-CHB, SRP-PHAT, MUSIC methods, and the discrete estimates of the EB-ESPRIT method, respectively, over SNR and T_{60}. For the cases of static speakers, the EB-ESPRIT uses the overall average (4 seconds) of segmental covariance matrices to achieve best accuracy. It has discrete DOA estimates which are plotted in the diamond symbol on the horizontal axes. The ground truth DOA is 67°.

From Fig. 3.9, the proposed Onset-GSRP produces consistent peaks at around 67° over T_{60} from 0.2s to 1s and SNR from ∞ to 10dB. The SRP-PHAT and MUSIC also consistently produce peaks at around the ground truth DOA. The TF-CHB however, shows considerable deviations due to the reverberation or noise. The EB-ESPRIT produces DOA estimates close to the ground truth at $T_{60} = 0.2$s, but otherwise shows significant errors due to reverberation. It is interesting to note that the EB-ESPRIT is robust against noise, which is expected from its formulation.

Fig. 3.10 shows estimation results from high resolution estimators. Clearly both the proposed Onset-MCC and MCC-PHAT produce accurate peaks corresponding to the ground truth, except for the spurious peaks at $T_{60} = 1$s and SNR=10dB. The Neuro-Fuzzy method produces more spurious peaks,

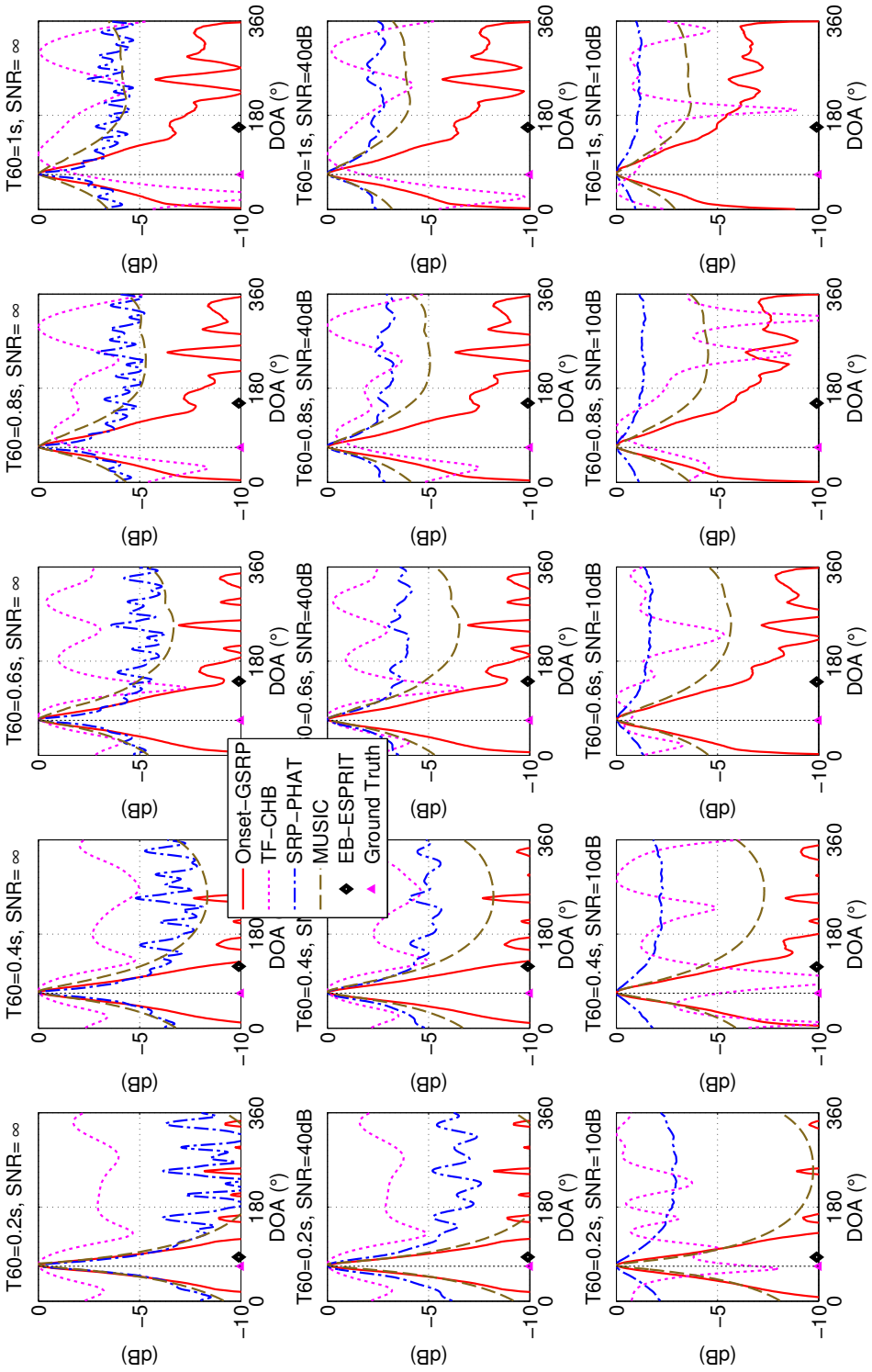

Figure 3.9: One static source located not close to the wall, at DOA of 67°.

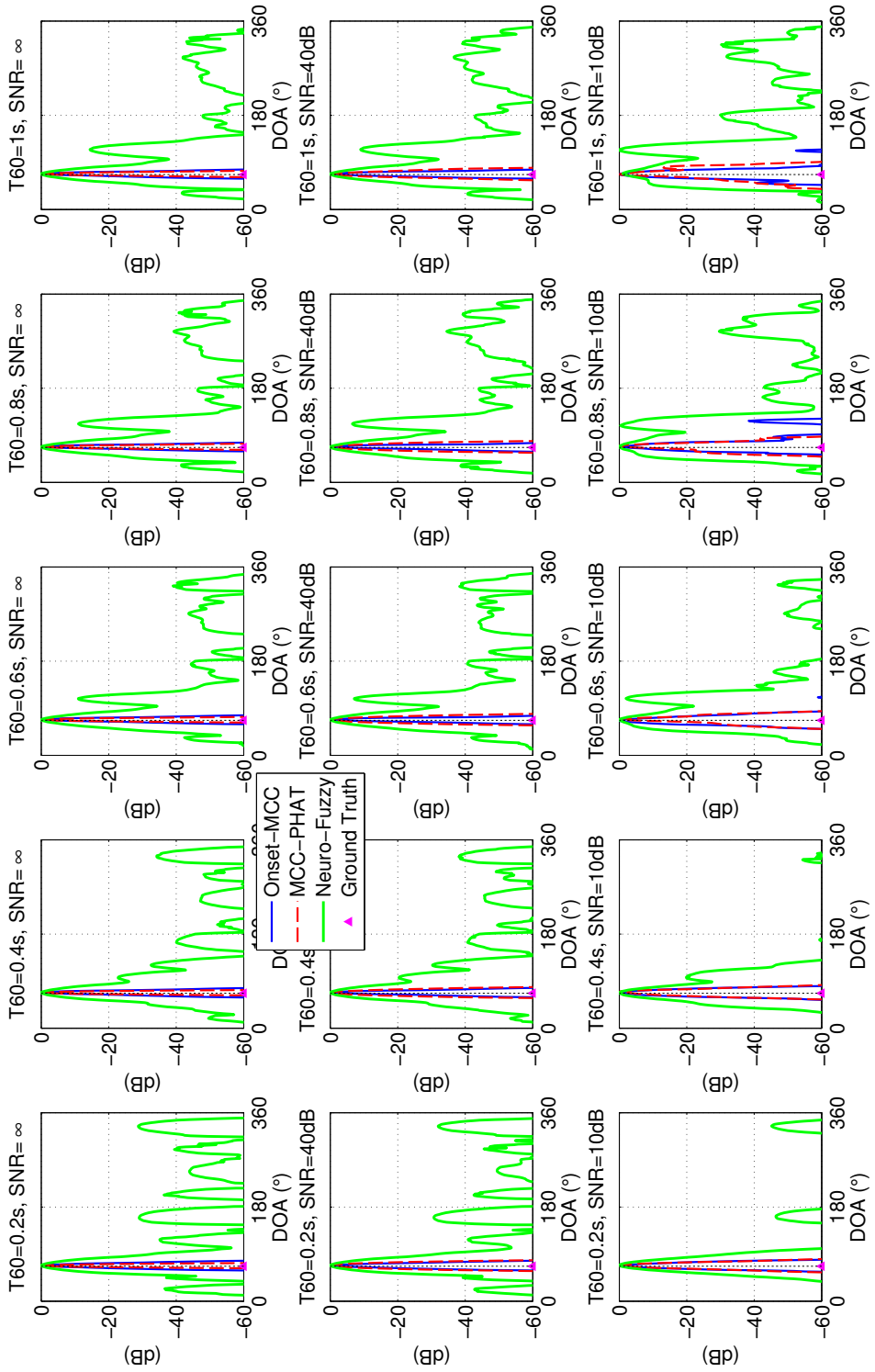

Figure 3.10: One static source located not close to the wall, at DOA of 67°.

Table 3.1: RMSE of DOA estimation results (in degrees) using different methods. DOA=67°.

	SNR = ∞					SNR = 40dB				
Methods	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1
Onset-GSRP	0	0	0	0	1	0	0	0	1	1
Onset-MCC	0	0	0	0	0	0	0	0	0	0
MCC-PHAT	0	0	0	0	0	0	0	0	0	1
Neuro-Fuzzy	0	0	0	0	0	0	0	0	0	0
TF-CHB	2	4	8	230	37	3	5	7	231	37
EB-ESPRIT	16.49	50.26	73.62	83.67	89.22	16.49	50.23	73.58	83.67	89.23
SRP-PHAT	0	0	0	0	0	0	0	1	1	1
MUSIC	3	2	1	0	0	2	2	1	0	0
	SNR = 20dB					SNR = 10dB				
Methods	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1
Onset-GSRP	1	0	1	1	1	1	0	1	1	2
Onset-MCC	0	0	0	0	1	0	1	2	0	0
MCC-PHAT	1	0	1	1	1	0	2	1	2	1
Neuro-Fuzzy	1	0	1	44	47	1	0	0	0	46
TF-CHB	9	225	227	23	4	253	218	43	36	30
EB-ESPRIT	16.42	50.02	73.26	83.62	89.29	16.29	49.49	72.52	83.50	89.46
SRP-PHAT	0	0	1	2	2	1	1	1	3	3
MUSIC	1	1	1	0	0	1	1	1	1	1

especially at highly reverberant conditions.

The RMSE of the DOA estimates are provided in Table 3.1. Note that as already pointed out in the experimental setup, the azimuthal DOAs are scanned in 1° steps. Thus the RMSE becomes 0° for the cases when the deviation is small (within 1°) and the correct DOA (i.e. 67°) is obtained in all frames (i.e. all the peaks of localization functions correspond to 67°). It is clear that using the 1° step is sufficient to characterize and compare the performance of all studied methods. Choosing a DOA resolution of smaller than 1° is neither necessary nor too useful for speaker localization in practice.

SCENARIO 2 - A STATIC SOURCE CLOSE TO WALL (SIMULATION)

Fig. 3.11 and Fig. 3.12 provide the histograms of localization functions from the proposed Onset-GSRP, Onset-MCC and MCC-PHAT methods as well as that of the Neuro-Fuzzy method, the steered-response power of the TF-CHB, SRP-PHAT, MUSIC methods, and the discrete estimates of the EB-ESPRIT method, respectively, over SNR and T_{60}. The static source locates at 0.2m from the wall (cf. Fig. 3.3), at the DOA of 45°. In this case, the early reflection from the closest wall is about 1ms behind the direct-path. With additive noise at SNRs from ∞ to 10dB, it is interesting to see that all the methods work fine at $T_{60} = 0.2$s, while as the reverberation increases, the EB-ESPRIT could not find the correct DOAs (even though given the *a priori* knowledge of one active source), and the dominant peak SRP of the TF-CHB also deviates significantly from 45°.

The Onset-GSRP, SRP-PHAT and MUSIC all produce correct peaks in all the cases of this scenario. In Scenario 1 and 2 (cf. Fig. 3.4), the shortest distance from sources to wall is about 1m, which corresponds to about 6ms time delay between the direct-path and the first early reflection. This follows the assumptions and discussions about the early reflections of the RIR model in Section 2.3.2. When the source is located close to (within about 1m) reflective surfaces, the performance of these localization methods may degrade, but in general remains close to that of Scenario 1. Thus the onset detection and encoding methods work reliably for the Onset-GSRP and Onset-MCC.

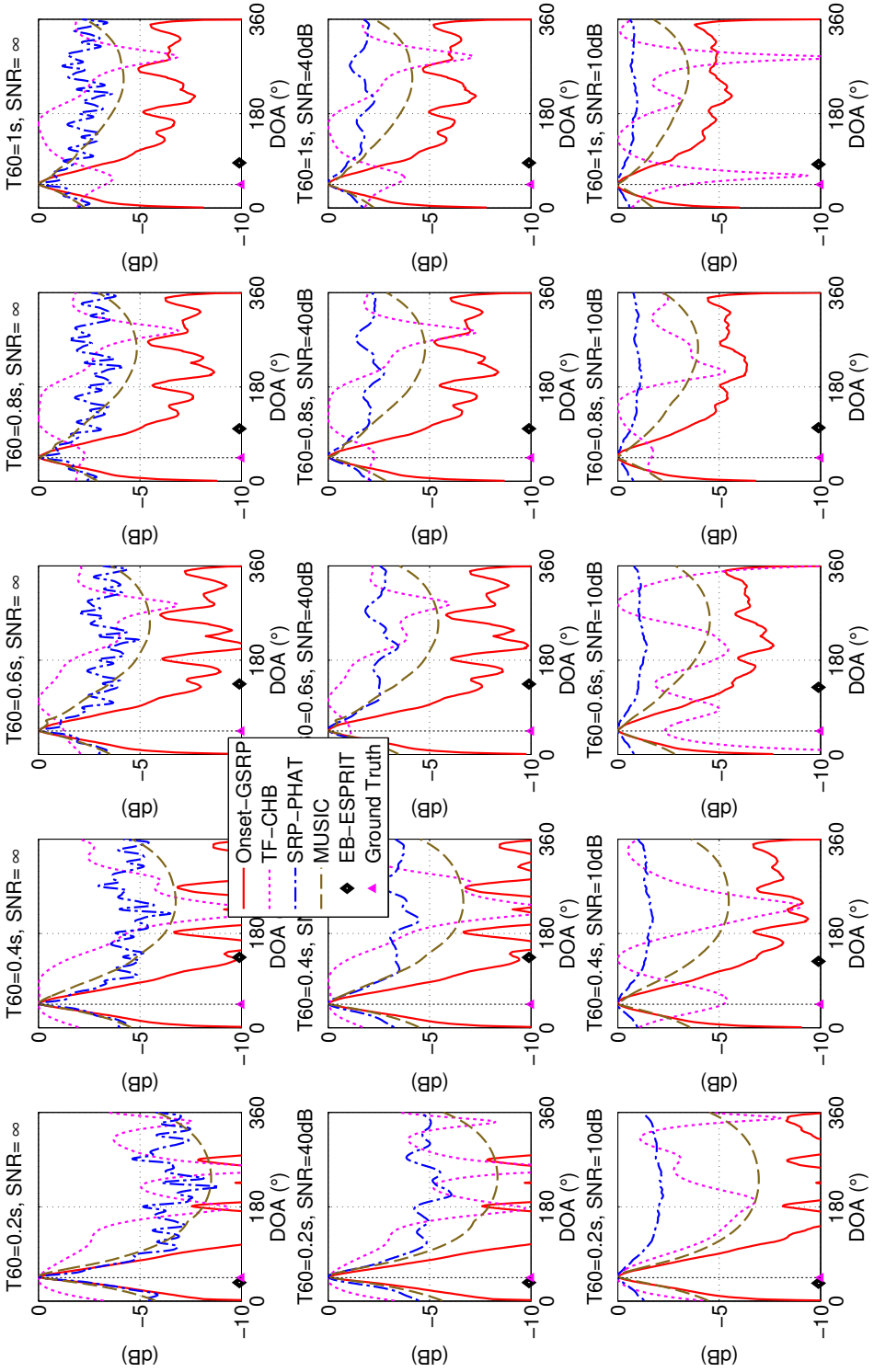

Figure 3.11: One static source located at 0.2m from the wall, at DOA of 45°.

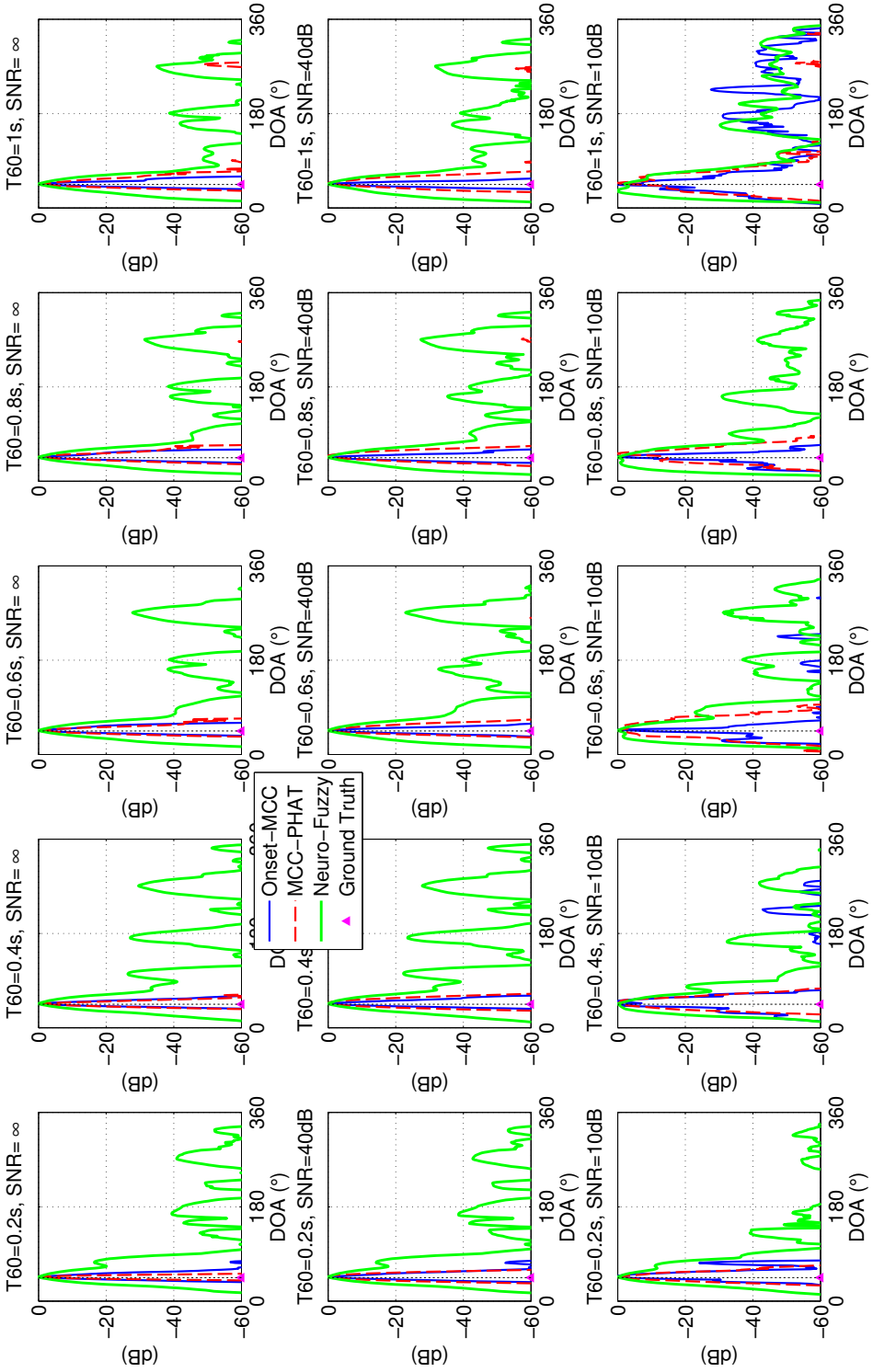

Figure 3.12: One static source located at 0.2m from the wall, at DOA of 45°.

Table 3.2: RMSE of DOA estimation results (in degrees) using different methods. DOA=45°, close to wall.

	SNR = ∞					SNR = 40dB				
Methods	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1
Onset-GSRP	0	0	0	0	0	0	0	0	0	0
Onset-MCC	0	0	0	0	0	0	0	0	0	0
MCC-PHAT	0	0	0	1	0	1	1	0	3	2
Neuro-Fuzzy	0	0	0	0	0	0	0	0	0	0
TF-CHB	5	1	56	74	113	5	43	56	76	112
EB-ESPRIT	10.45	88.50	88.82	54.38	40.29	10.49	88.27	88.62	54.45	40.22
SRP-PHAT	0	0	0	0	0	1	2	3	4	3
MUSIC	2	0	0	0	0	1	0	0	0	0
	SNR = 20dB					SNR = 10dB				
Methods	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1
Onset-GSRP	0	0	0	0	0	1	0	0	1	1
Onset-MCC	1	0	0	0	1	2	5	1	1	1
MCC-PHAT	1	2	3	7	2	0	1	0	7	2
Neuro-Fuzzy	0	0	0	0	0	0	0	0	0	13
TF-CHB	3	77	4	94	95	2	85	243	96	86
EB-ESPRIT	10.87	86.17	86.84	55.05	39.55	11.77	80.86	82.47	56.33	37.48
SRP-PHAT	2	2	3	5	2	3	4	3	7	2
MUSIC	1	0	0	0	1	2	0	1	0	1

In Fig. 3.12, while the strongest peaks of the Onset-MCC and MCC-PHAT all correspond to the true speaker DOA, the Neuro-Fuzzy at SNR=10dB and $T_{60} \geq 0.8$s does not. The Onset-MCC produces spurious peaks at SNR=10dB, and Neuro-Fuzzy has more spurious peaks in the given range. The peaks of these three high resolution methods grow only slightly wider as the reverberation and noise increases. The RMSE of the DOA estimates are provided in Table 3.2.

SCENARIO 3 - TWO STATIC SPEAKERS (SIMULATION)

Fig. 3.13 and Fig. 3.14 plot the DOA localization results for the case (cf. Fig. 3.4) when two static speakers locate at DOAs of 170° and 190°, respectively. Overall, from Fig. 3.13, the TF-CHB forms a wide peak at around 180° in most cases except $T_{60} \geq 0.8$s, indicating that the two speakers are also fused. The EB-ESPRIT again assumes a known number of speakers (which avoids the errors due to estimation of the number of speakers at adverse conditions), and produces a DOA estimate at around 180° and a second estimate at close to 0°. This indicates that the EB-ESPRIT also has ambiguity to differentiate the two speakers. Except at $T_{60} = 0.2$s and high SNR, the Onset-GSRP, the SRP-PHAT and MUSIC also fuse the two sources into one (cf. Appendix F and Appendix G). In Fig. 3.14, it is obvious that the Neuro-Fuzzy method no longer forms two distinct peaks and the two speakers are fused into one speaker in the estimation. However, the proposed Onset-MCC and MCC-PHAT methods can reliably form two distinct peaks corresponding to the ground truth in most cases.

Fig. 3.15 and Fig. 3.16 plot the DOA localization results for the case (cf. Fig. 3.4) when two static speakers locate at DOAs of 165° and 195°, respectively. From Fig. 3.15, the TF-CHB and EB-ESPRIT still fuse the two speakers into one. MUSIC can also resolve the two speakers in most cases. The SRP-PHAT can only resolve the speakers at low reverberation and high SNR cases. The proposed Onset-GSRP can resolve the speakers for cases of $T_{60} \leq 0.8$s when SNR$\geq$40dB. In Fig. 3.16, all the three high resolution methods can reliably form two distinct peaks corresponding to the ground truth in most

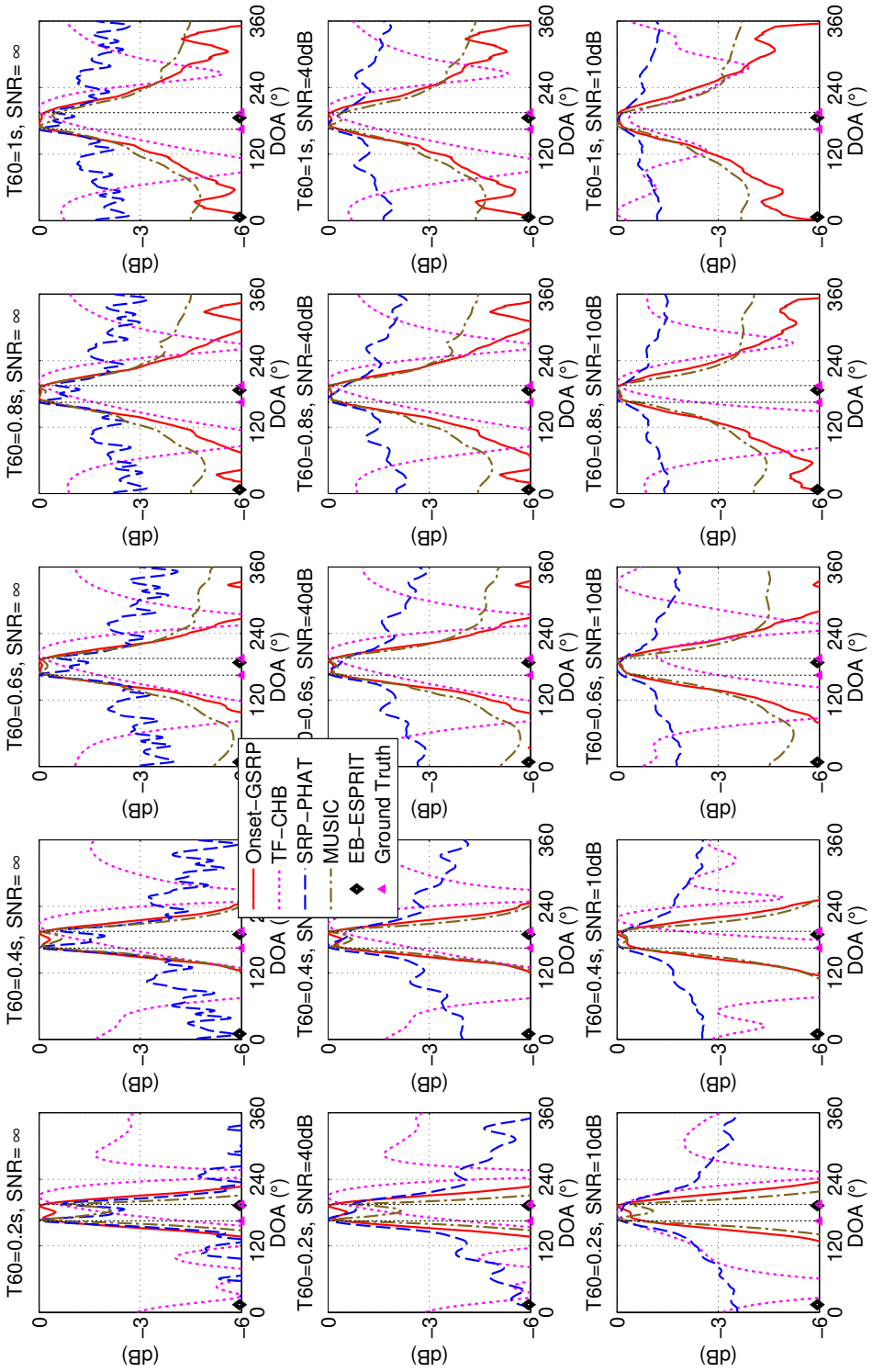

Figure 3.13: Resolution study of different methods at levels of reverberation and noise. Ground truth DOAs are marked as triangles at 170° and 190°.

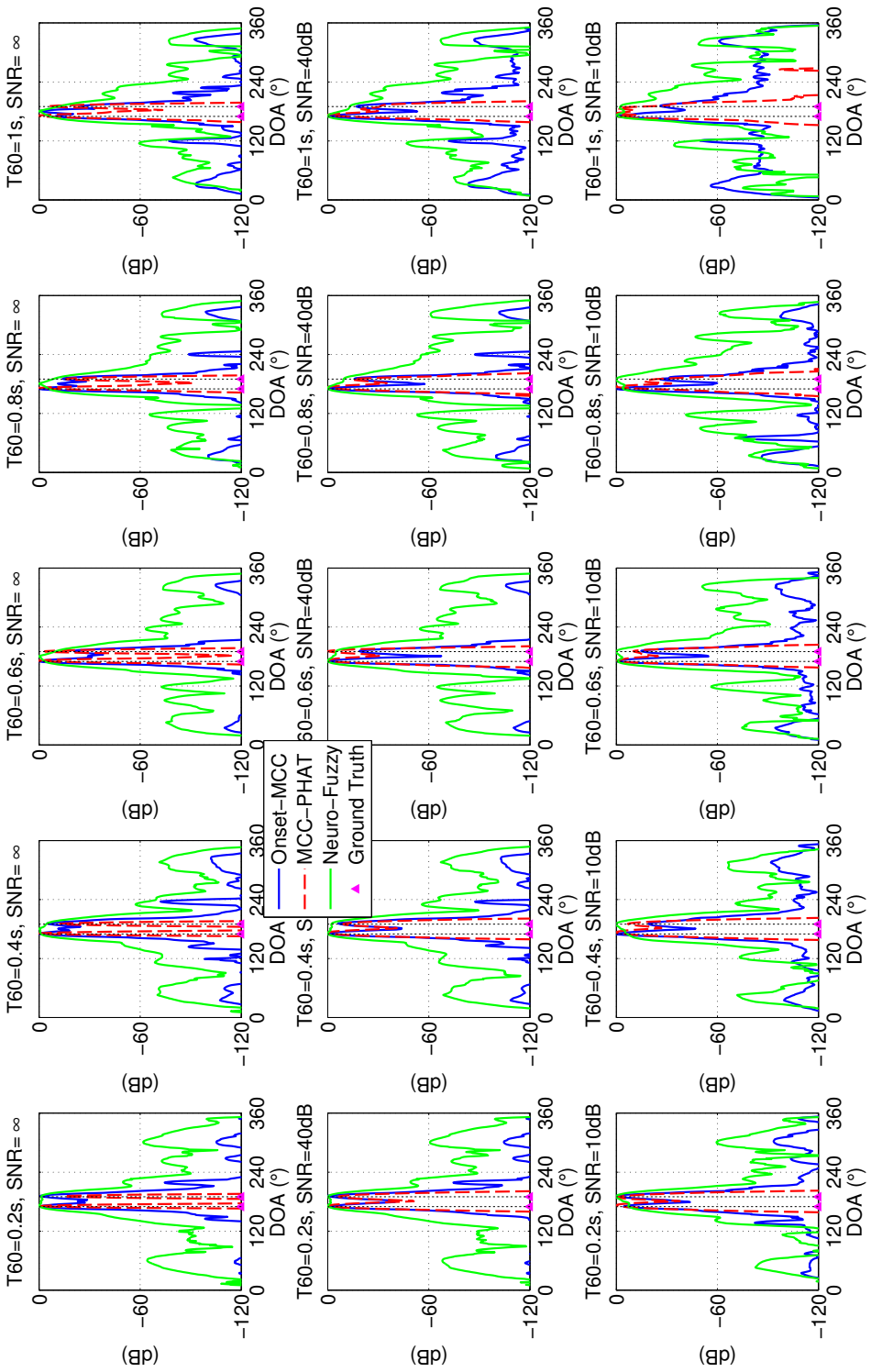

Figure 3.14: Resolution study of different methods at levels of reverberation and noise. Ground truth DOAs are marked as triangles at 170° and 190°.

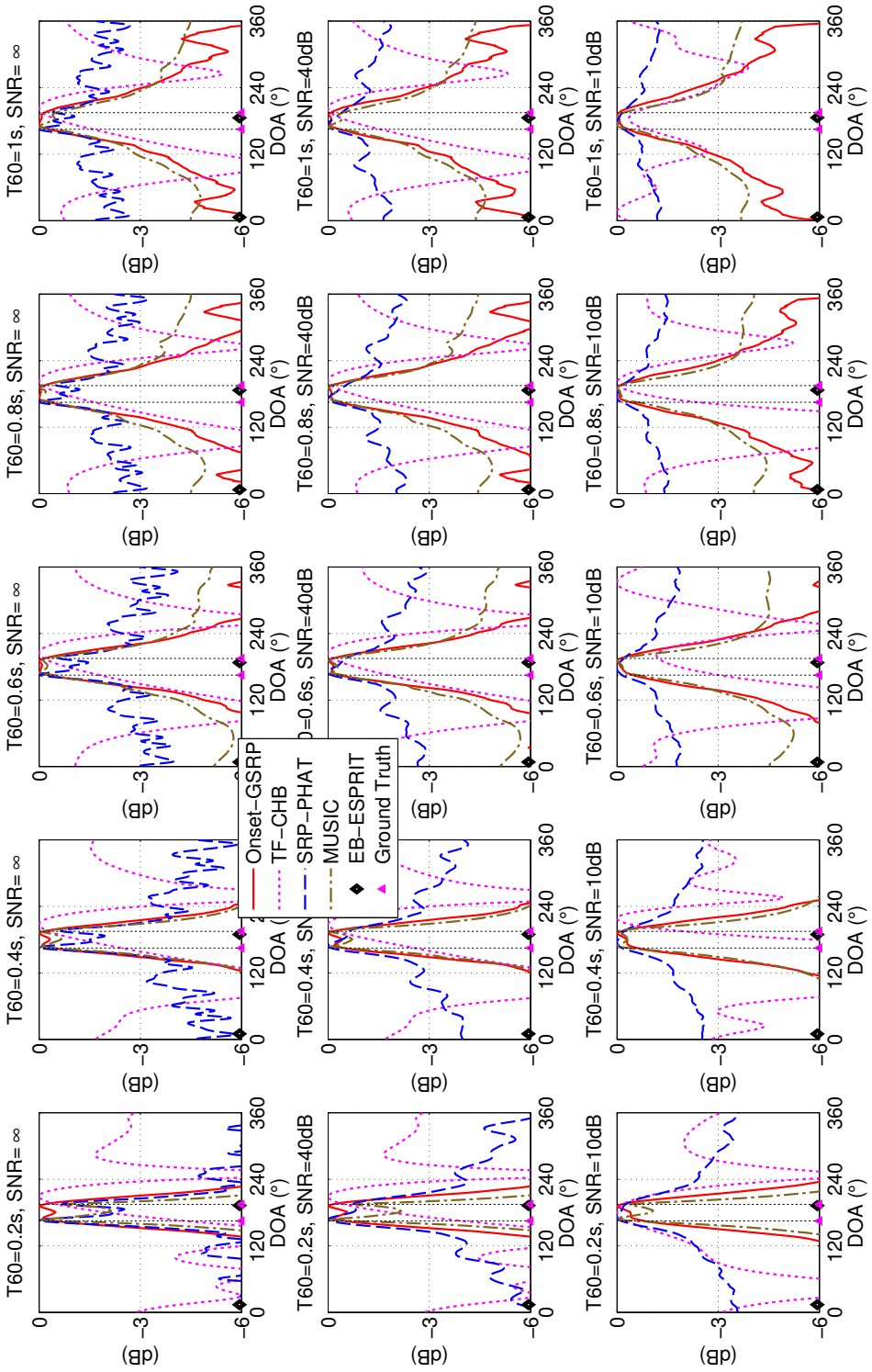

Figure 3.15: Resolution study of different methods at levels of reverberation and noise. Ground truth DOAs are marked as triangles at 165° and 195°.

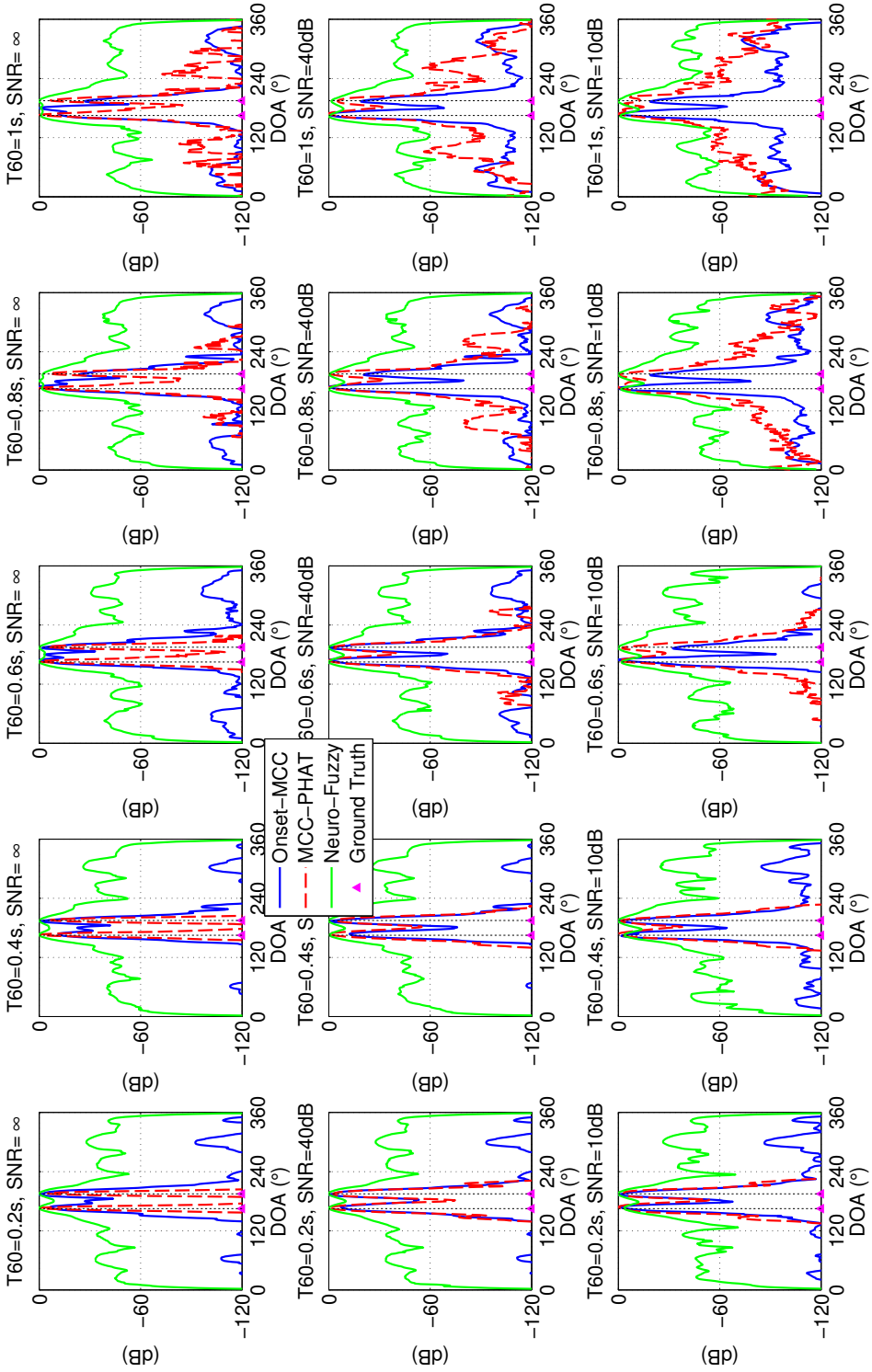

Figure 3.16: Resolution study of different methods at levels of reverberation and noise. Ground truth DOAs are marked as triangles at 165° and 195°.

cases.

The RMSE of the DOA estimates are provided in Table 3.3. For those cases when estimators do not produce the right number of estimates reasonably close to speaker DOAs, quantitative accuracy measures are not provided.

SCENARIO 4 - THREE STATIC SOURCES (SIMULATION)

Fig. 3.17 and Fig. 3.18 provide localization results for three static speakers. Here it is assumed that the number of speakers is known *a priori* to the EB-ESPRIT.

From Fig. 3.17, it can be seen that the SRP peaks from the TF-CHB are close to ground truth DOAs at low reverberation ($T_{60} = 0.2$s), but are relatively wider than other methods and much disturbed as the reverberation gets stronger. The DOA estimates from the EB-ESPRIT can be accurate, but show higher offsets at $T_{60} = 0.4$s and $T_{60} = 1$s than other cases. It can also be seen that the EB-ESPRIT is robust against white Gaussian noise as this matches the underlying noise subspace model it assumes. Note here that assuming a known number of speakers gives EB-ESPRIT a considerable advantage, as this avoids the errors due to the estimation of the number of speakers. Both the well accepted SRP-PHAT and MUSIC, as well as the proposed Onset-GSRP consistently produce three peaks corresponding to the ground truth DOAs. Peaks of the MUSIC are most distinct at low reverberation, but get wider as the reverberation increases. Peaks of the SRP-PHAT are most distinct at no noise cases (except for $T_{60} = 0.2$s), but get considerably wider as the SNR drops. In comparison, the peaks of Onset-GSRP are relatively the most consistent, indicating robustness against reverberation and noise for the case of three concurrent speakers. The ranges of SRPs from these methods are relatively confined, and hence are plotted in the same figure.

Fig. 3.18 plots the results from high resolution location estimators, i.e. the Onset-MCC, MCC-PHAT and the Neuro-Fuzzy. The MCC-PHAT has the best resolution (most distinct peaks) overall, and for most cases can produce peaks corresponding to the ground truth DOAs. The Neuro-Fuzzy method

Table 3.3: RMSE of DOA estimation results (in degrees) using different methods. Two sources closely located.

{170°, 190°}	SNR = ∞					SNR = 40dB				
Methods	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1
Onset-GSRP	-	-	-	-	-	-	-	-	-	-
Onset-MCC	1	1	1	0.71	-	1	1	0.71	0.71	1
MCC-PHAT	0	0	0.71	1	0.71	1	1.58	0.71	2	-
Neuro-Fuzzy	0	-	-	-	-	0	2.12	2.12	-	-
TF-CHB	-	-	-	-	-	-	-	-	-	-
EB-ESPRIT	-	-	-	-	-	-	-	-	-	-
SRP-PHAT	0.71	1	-	-	-	-	-	-	-	-
MUSIC	1.58	1.58	-	-	-	1	1.58	-	-	-
{165°, 195° }	SNR = ∞					SNR = 40dB				
Methods	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1
Onset-GSRP	4	4	4	6.96	6.96	4	4	4	6.96	6.96
Onset-MCC	3	2.55	3	2.55	13.51	3	2.55	3	2.55	2.55
MCC-PHAT	0	0	0	0	0	1	1	1.58	0.71	2.55
Neuro-Fuzzy	0	0.71	1.41	23.72	2	0	0.71	1.58	2	2.55
TF-CHB	-	-	-	-	-	-	-	-	-	-
EB-ESPRIT	-	-	-	-	-	-	-	-	-	-
SRP-PHAT	0	0	0	0.71	1	2.55	4.74	-	-	-
MUSIC	1	1.58	3	5.83	4.74	1	1.58	3	2.55	4.74

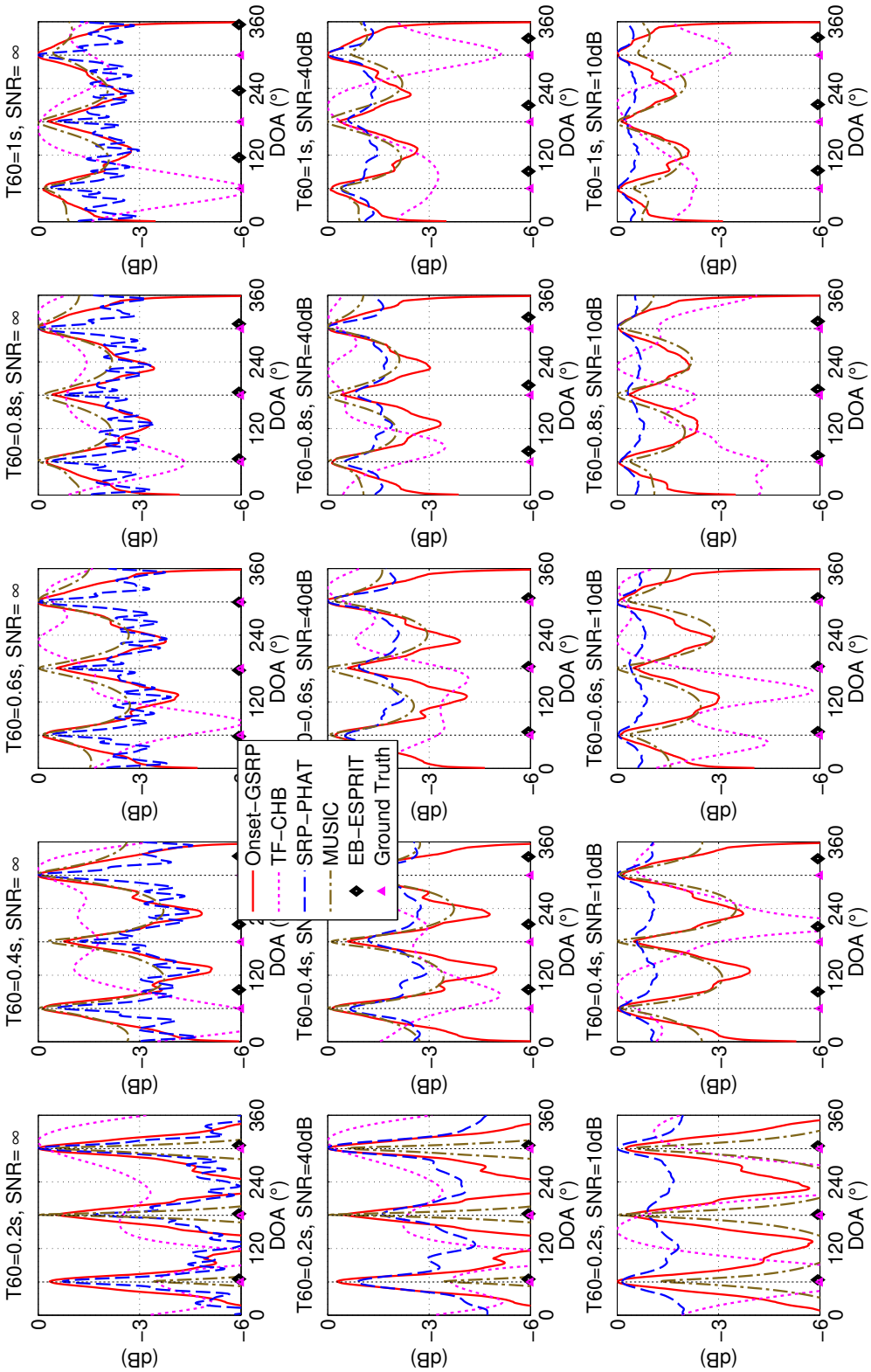

Figure 3.17: Normalized histograms from the Onset-GSRP, SRP-PHAT and MUSIC localization functions, steered-response power from the TF-CHB method and DOA estimates from the EB-ESPRIT method. Triangles mark the ground truth DOAs.

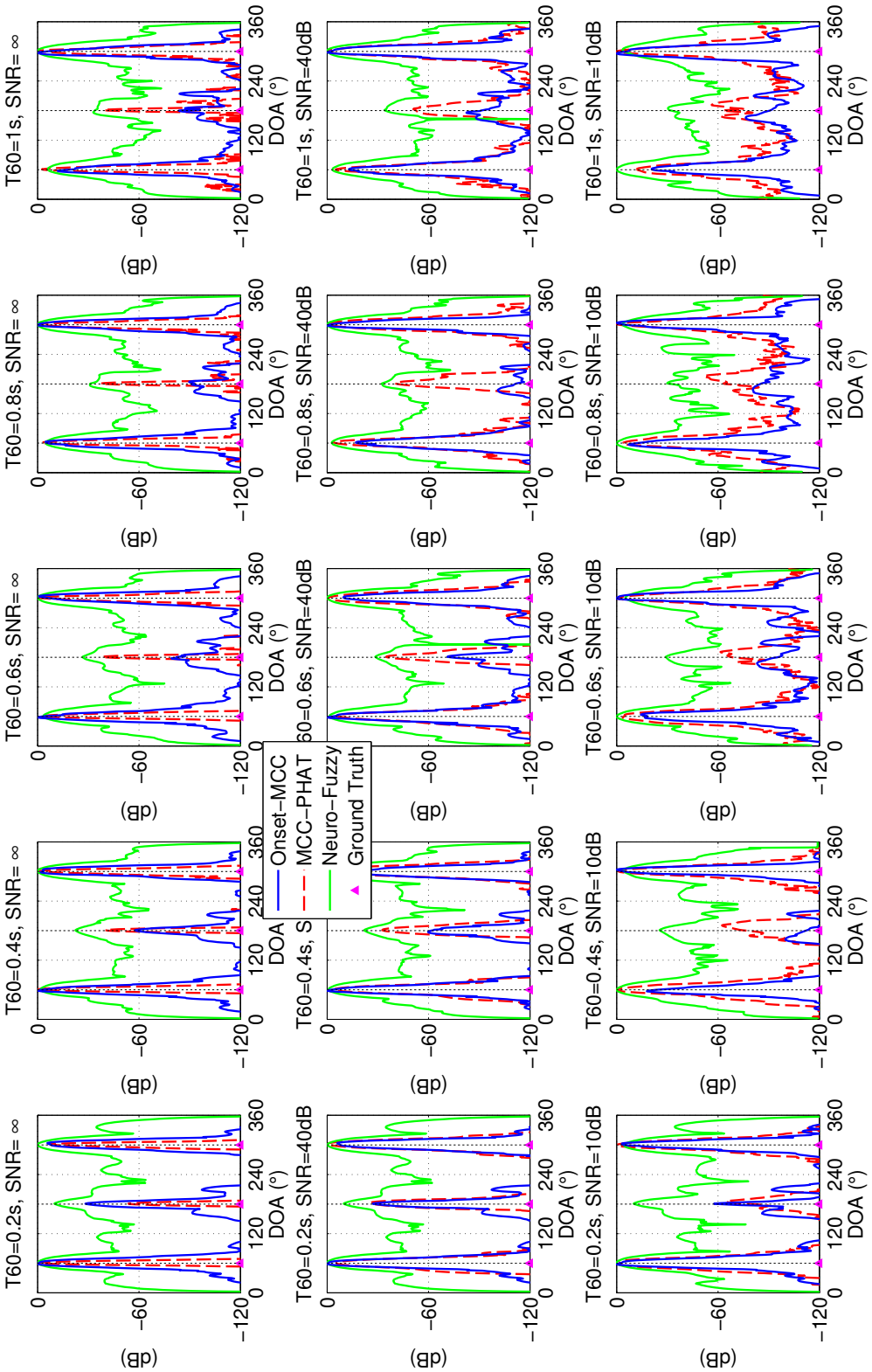

Figure 3.18: Normalized histograms from the Onset-MCC, MCC-PHAT and Neuro-Fuzzy localization functions. Triangles mark the ground truth DOAs.

Table 3.4: RMSE of DOA estimation results (in degrees) using different methods. Three concurrent static speakers.

	SNR = ∞					SNR = 40dB				
Methods	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1	$T_{60}(s) = 0.2$	0.4	0.6	0.8	1
Onset-GSRP	0	0.58	1.73	1.73	2.08	0	0.58	1.73	2.45	1.83
Onset-MCC	2.08	2.08	3	1.91	2.51	2.08	3.37	1.83	7.59	3.37
MCC-PHAT	0	0	0	0	0	1.29	0	0.82	1	1.41
Neuro-Fuzzy	0.58	0	0	1.29	2.89	0.58	0	0.58	0	0
TF-CHB	11.11	49.74	57.74	–	–	12.37	–	61.89	40.68	–
EB-ESPRIT	3.93	32.27	3.14	5.52	54.63	3.90	32.08	4.83	18.32	29.30
SRP-PHAT	0	0	0.82	0.82	0.82	1	1.91	1.73	1.73	1.91
MUSIC	0	0.58	0.82	1.41	1.63	0	0.58	1.15	1.63	1.15

produces wider peaks compared to the MCC-PHAT. However, the peaks corresponding to Speaker1 (male at 180°) get much weaker, especially as the noise level increases (SNR=10dB). This is easy to understand, as the encoding of (3.34) leads to superior DOA resolutions, but strong noises can disturb the peaks of subband signals hence the resulting cross-correlation coefficients in DOA estimation (cf. the discussion in Section 3.3.6). This indicates a trade-off between resolution and noise-robustness as discussed.

Table 3.4 provides the RMSE of the DOA estimates in this scenario from different methods for SNR $= \infty$ and 40dB. It is obvious that the Onset-GSRP, MCC-PHAT, Neuro-Fuzzy, SRP-PHAT and MUSIC methods achieve similar DOA estimation accuracies, which are within 3°. The Onset-MCC has slightly larger errors. The TF-CHB and EB-ESPRIT methods have overall larger errors compared to other methods. In general, the accuracy degrades as the reverberation increases. Note that TF-CHB does not produce three peaks in some cases (e.g. $T_{60} = 0.8$s and $T_{60} = 1$s), hence do not have valid RMSE results to present. Similarly for SNR=10dB, when there are miss-detections or spurious estimates, the RMSE measure may no longer be consistent or informative, due to the difficulty in mapping the estimates with the ground truth. Quantitative accuracy measure using the OSPA metric for such cases will be given in the next Scenario.

SCENARIO 5 - THREE MOVING SPEAKERS (REAL-WORLD)

In this case, the localizations of three moving speakers are estimated, in a real reverberant room (with measured $T_{60} \approx 0.65$s). The trajectories of speakers are given in Fig. 3.6. The raw speech signals of the three speakers are given in Fig. 3.7.

As noticed in Scenarios 3 and 4, it is neither straightforward nor informative to use the RMSE as the localization performance measure, when there are spurious estimates or miss-detections, especially when there are multiple moving speakers as the locations change over time. Therefore, the OSPA metric [122] is used here as a consistent localization performance measure. It takes into account the permutations of speakers and evaluates not only

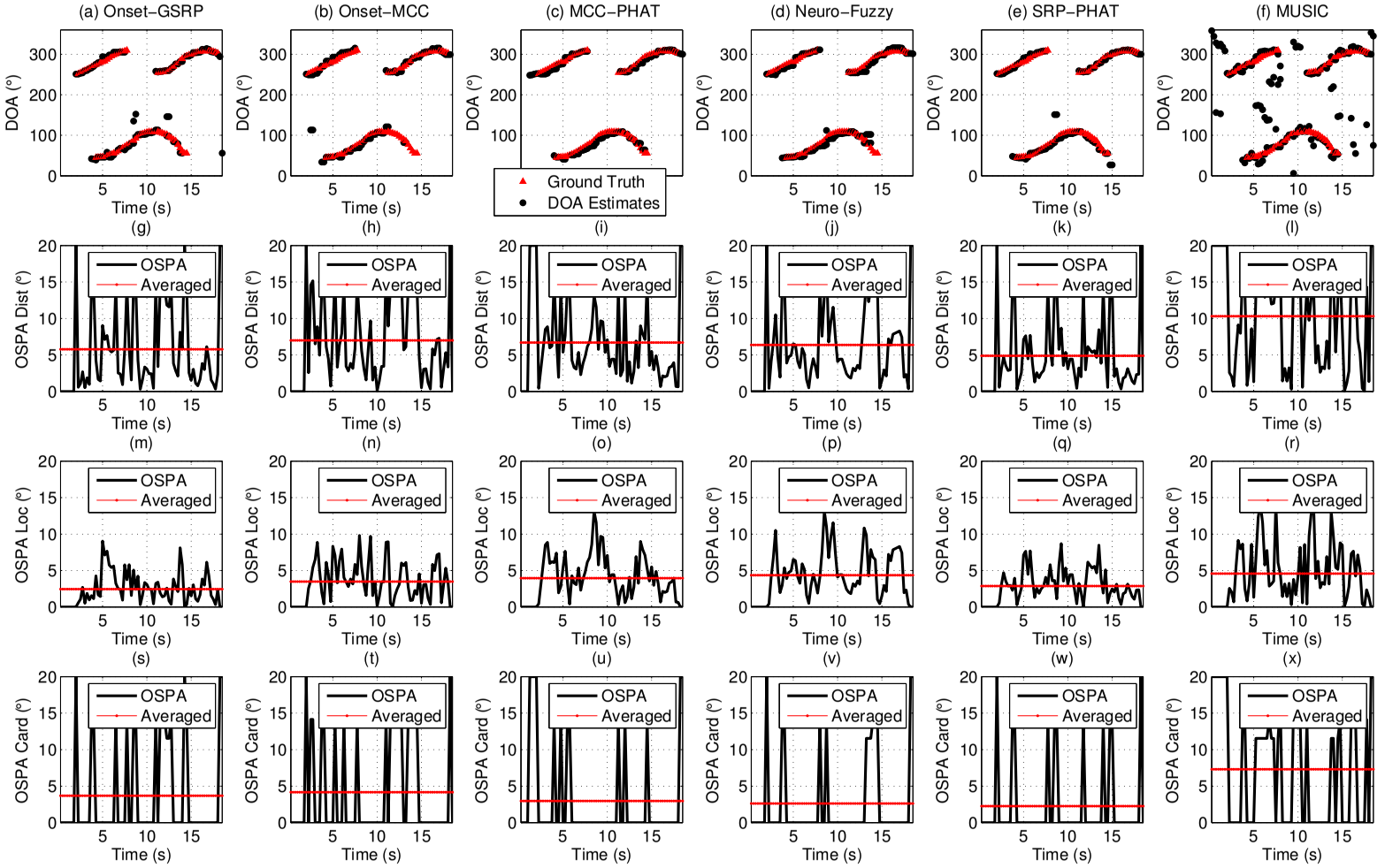

Figure 3.19: DOA estimates (top row) and OSPA results from different methods. Three speakers are moving and speaking in the real reverberant room ($T_{60} \approx 0.65$s).

the DOA miss-distances (offsets) but also the cardinality errors (errors in the estimated number of speakers).

The OSPA metric is defined as (2.113) in Section 2.6.6. Apparently when choosing $p = 2$, the OSPA metric can be viewed as an extended RMSE measure that selects closest estimate-truth pairs and includes the "penalty" for estimation errors in the number of speakers. Usually the parameter c is chosen to be greater than the maximum estimation deviation of respective localization methods. Here $p = 2$ and $c = 20^\circ$ are chosen.

The final DOA estimates from (3.62) and OSPA results are given in Fig. 3.19. Each column shows the results from a particular method, i.e. from the left column to the right column, the methods used are respectively the Onset-GSRP, Onset-MCC, MCC-PHAT, Neuro-Fuzzy, SRP-PHAT and MUSIC. Note that the Onset-GSRP has lower DOA resolution, hence requiring a larger θ_r (here 45° is used for Onset-GSRP, and 30° is used for the rest) to suppress spurious peaks in low noise cases. The TF-CHB and EB-ESPRIT do not produce reliable results in this case of moving speakers, and hence are not plotted. In the top panel of each column, DOA estimates over time are plotted in black dots, while the ground truth locations of speakers are plotted in red triangles. It can be seen from panels (a) to (e) that although there are several spurious peaks and gaps, they produce close and clean DOA estimates in general. The SRP-PHAT seems to have the best overall OSPA results. MUSIC has a significant amount of spurious estimates and the worst overall OSPA results, even though the ground truth number of speakers per each frame is provided for its localization function. This follows the discussions in Section 2.5.4 that subspace based methods may not work well for moving speakers with reverberation. In general, the Onset-GSRP, Onset-MCC, MCC-PHAT, Neuro-Fuzzy and SRP-PHAT methods demonstrate similar performance in this case of moving speakers with strong reverberation.

The OSPA results provide closer details of the DOA estimation errors. Panels (g) to (l) provide the overall OSPA distances, which are composed of the respective OSPA location errors (panels (m) to (r)) and OSPA cardinality errors (panels (s) to (x)), as shown in (2.113). Here the OSPA location errors measure the deviations from the location estimates to the ground truth

locations. It can be seen from (m) and (q) that the proposed Onset-GSRP method and the SRP-PHAT method have lowest averaged OSPA location error of about 3°, and the maximum errors over time is less than 10°. The MCC-PHAT and the Neuro-Fuzzy methods as in (o) and (p) produce slightly higher averaged OSPA location errors of about 4°, and the maximum errors over time is about 13°. MUSIC in (r) shows higher offsets than the other methods. The OSPA cardinality errors weigh the number of miss-detections and spurious estimates in the estimated DOAs. For example, in panel (s) at time of about 2s, a spurious estimate when there is no speaker gives an OSPA cardinality error of $c = 20^\circ$, while in panel (t) at about 5s, a miss-detection (only one speaker is detected) when there are two speakers gives an OSPA cardinality error of 14.1°. The averaged OSPA cardinality errors for the Onset-MCC, MCC-PHAT, Neuro-Fuzzy and SRP-PHAT methods are all around 3°. MUSIC has highest cardinality errors as it produces more spurious peaks due to moving speakers and reverberation. To sum up, the overall OSPA distances as shown in (g) to (l) demonstrate that the proposed Onset-MCC and MCC-PHAT methods as well as the Neuro-Fuzzy and SRP-PHAT can locate moving speakers with an averaged OSPA error of about 6°, while the MUSIC method produces considerably larger localization errors.

3.8 Conclusions and Discussions

3.8.1 Conclusions

This chapter proposes three novel reverberation-robust speaker localization algorithms, which are referred to as the Onset-MCC, Onset-GSRP and MCC-PHAT, respectively. The Onset-MCC and Onset-GSRP algorithms first decompose speech mixtures into subbands via an auditory filterbank, based on the speech signal model and the TF sparsity assumption. Then a novel onset detection and encoding approach is derived to extract the direct-path components from reverberant microphone recordings, based on the speech signal and the acoustic RIR models. Furthermore, the subband cross-correlation coefficients of the direct-paths signals are reversely mapped from relative sam-

ple delays to locations, and hence produce overall DOA localization functions. The MCC-PHAT method builds upon the classic GCC-PHAT method, and exploits the redundant information from multiple closely placed microphones to suppress the impact of reverberation.

Performance of the presented methods is studied using not only simulated signals of reverberation time from 0.2 to 1s, but also real recordings in an office room of $T_{60} \approx 0.65$s. Comparison with other baseline localization techniques in various reverberant conditions demonstrates that the proposed Onset-GSRP, Onset-MCC and MCC-PHAT localization algorithms can reliably locate not only static speakers but also multiple moving speakers, in presence of strong reverberation. The proposed Onset-MCC and MCC-PHAT methods achieve better DOA resolutions compared with all other benchmark methods. Appendix K also provides an analysis of the computational complexities of the localization algorithms.

3.8.2 Discussions

Part I, in summary, studies the problem of speaker location (DOA) estimation *in short time frames*, acoustically. The speaker DOA is estimated in each time frame, while the available data are the snapshots of sound signals captured by microphone arrays. Although encouraging results have been obtained in comparison with the state-of-the-art methods, the DOA estimates may still contain spurious estimates or miss-detections in challenging conditions, and the association with speakers and between DOA estimates at consecutive time frames remains unclear, especially for moving speakers.

Therefore, Part II of the thesis explores the problem of tracking speaker states *over time* using imperfect observations (e.g. location estimates from Part I). This requires associating correct identities to respective speaker states, addressing the spurious estimates or miss-detections in the feature estimates, and resolving the ambiguity when the single feature estimates of different speakers are close. Thus Part II proposes two methods for speaker feature filtering, based on the GLMB Bayes RFS framework. It first develops in Chapter 4 an adaptive multi-speaker tracking filter that produces labeled trajecto-

ries of speakers over time, which is evaluated in a reverberant environment. Then in Chapter 5, by generalizing the speaker state to include not only the kinematic feature, but also pitches and sound waveforms, a multi-feature multi-speaker state filter is proposed, which can jointly track and separate multiple features from multiple speakers, and help to resolve the ambiguity problem that arises in the single feature filtering.

Part II

Feature Filtering

Chapter 4

Adaptive Bayes RFS Multi-speaker Tracking

This chapter investigates the adaptive filtering of the kinematic feature of speakers, in the challenging reverberant scenario. As discussed in Part I, numerous speaker localization methods can be found in the literature, including the subspace based methods [27–30], steered response power beamformers [10, 31–34], and TDOA based methods [35–39]. Most of the localization methods may degrade or break down under challenging conditions. Consequently, although the localization alone may provide location candidates of speakers at discrete time frames, there however can be gaps or miss-detections of speaker locations over time due to the nonstationarity of speech signals and interference of concurrent speakers, and clutter or spurious estimates due to reverberation. Moreover, consistently obtaining speaker trajectories via filtering unordered location estimates and associating them with corresponding speakers is also a significant challenge. Therefore, this chapter presents the multi-speaker tracking filter based on the Onset-MCC and MCC-PHAT localization methods proposed in Part I, as well as the GLMB Bayes RFS framework (cf. Section 2.6.4).

4.1 Introduction

A number of multi-speaker tracking methods following the localization step [1,2,4,5,8] have been developed. Assuming in general the spurious peaks induced by reverberation exhibit no temporal consistency from one time frame to the next, while the peaks corresponding to speakers follow a kinematic model, a particle filtering (PF) method using GCC and SBF front-ends was implemented [2] to mitigate the reverberation problem. The idea is further generalized [4] under a Bayes RFS filter framework that estimates TDOA based on microphone pairs, and extracts tracks using an RFS bootstrap Se-

quential Monte Carlo (SMC) filter. The above methods however, were evaluated at modest or low reverberation time ($T_{60} \leq 0.39$s). Another particle filter based algorithm has been developed in [142] using mutual information (MI) and voice activity detection (VAD) measures. Its performance was evaluated at reverberation time of $T_{60} = 0.35$s. It is unclear however how the algorithms perform at more reverberant environments and when competing speakers are moving while talking. Moreover, these tracking methods do not systematically provide identities of trajectories, while in practice it is useful to associate the trajectories to respective speakers. Recently a multi-speaker tracker was developed [8], which adopts a computational auditory scene analysis (CASA) [46] approach for localization and a "nearest neighbour" type of multi-speaker tracker, and includes a simulated test case of one moving speaker with concurrent static speakers. However, the "nearest neighbour" tracker requires a heuristic time-to-live constant to bridge gaps in static tracks, which may not suit complicated speaker motions, e.g. when concurrent speakers are moving.

Based on the localization results in Part I, this chapter formulates a new filter to estimate and track kinematic states of an unknown and time-varying number of moving acoustic sources in highly reverberant environments, from sound mixtures acquired by microphone arrays. The system consists of a reverberation-robust location estimator proposed in Part I and the GLMB Bayes RFS multi-object tracking framework. Uniform circular arrays (UCA) are used for estimating the source DOAs. The location estimator extracts the DOA measurements from sources via the MCC-PHAT or the Onset-MCC. Cartesian coordinates of sources are then tracked and labeled via the GLMB recursions. The MDB model is used for adaptive tracking of moving objects. The proposed framework has been evaluated using real recordings in a reverberant room ($T_{60} = 0.65$s).

This chapter is organized as follows. Section 4.2 shows an overview of the proposed system. The adaptive tracking filter is described in Section 4.3. Numerical studies of the proposed methods are presented in Section 4.4, and conclusions are given in Section 4.5.

4.2 System Overview

Bayes RFS filters (cf. Section 2.6) have been the emerging family of closed-form solutions to multi-object tracking [15–23]. The Probability Hypothesis Density (PHD) [15,16] and the Cardinalized PHD (CPHD) [17,19] have been well-accepted multi-object Bayes RFS tracking filters. Performance comparison of CPHD with traditional methods can be found in [19], which shows that CPHD is favourable than the PHD filter. However, CPHD still only propagates the first-moment (intensity) and cardinality distributions. The most current development of the Bayes RFS kind is the Generalized Labeled Multi-Bernoulli (GLMB) filter, which jointly tracks multi-object labeled states. Thus the multi-speaker tracking filter is developed based on the GLMB Bayes RFS framework [21–23]. For adaptive tracking, the measurement-driven object birth model for GLMB [25] is implemented, and due considerations are also given to the GLMB filtering of multi-sensor measurements.

As shown in Fig. 4.1, the proposed system consists of two stages in general, viz. the acoustic feature extraction (localization), and the multi-object Bayes RFS filtering (GLMB) as discussed in Section 2.6.4.

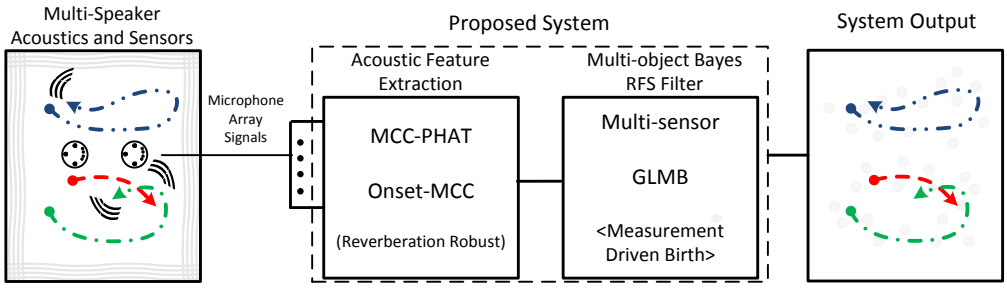

Figure 4.1: The diagram of the proposed multi-speaker tracking framework. Reverberated sound mixtures are acquired by circular Microphone Arrays. Location (DOA) estimates of speakers are obtained at the acoustic feature extraction stage based on the MCC-PHAT and Onset-MCC methods and further filtered by the multi-sensor MDB GLMB tracking framework. Resulting tracks of speakers are separated and labeled.

In the multi-speaker tracking implementation, the first stage is the acoustic source localization. The high resolution MCC-PHAT and Onset-MCC meth-

ods are used for reverberation-robust localization [45]. The estimated DOAs are then used to derive Cartesian locations based on locations of microphone arrays. The unordered location candidates may contain spurious estimates around true locations of speakers, as well as miss-detections. Moreover, tracking the kinematic states (e.g. trajectories) of respective speakers is also practically important, hence further filtering is necessary.

The second stage is the GLMB filter, which is supplied with the location candidates from multiple microphone arrays of the first stage. This multi-object tracking filter recursively processes the candidate locations, via a Bayesian prediction and update recursion, thereby providing the estimated speaker trajectories. The SMC implementation of GLMB is used because of the nonlinear motion of speakers in enclosed environments. The measurement-driven birth model for GLMB (cf. Section 2.6.4) is implemented [25] for adaptive tracking [23, 120]. The filter provides not only trajectories of locations but also the associated label (speaker identity) for each estimated track.

4.3 Filter Implementation

The GLMB recursion consists of the prediction and update steps. In the prediction step, each object transits to a new state according to the state transition function $f(\mathrm{s}|\cdot,\ell)$ in (2.104e). In the update step, the state is confirmed with the measurement based on the likelihood function $g(z_{\vartheta(\ell)}|\mathrm{s},\ell)$ in (2.102d). Since a nonlinear transform is required to convert DOA estimates to Cartesian coordinates, care must be taken for the likelihood function when the MDB model is used for adaptive filtering.

4.3.1 State Transition Function

The motion of speakers is rather random and unpredictable. Hence it would be highly impractical to assume complete knowledge of the transition function for speakers. For practical tractability however, it is often assumed that the motion of the speaker follows the Langevin model [1, 2, 4], which de-

scribes also a first-order Markov process. Thus for each dimension of the Cartesian coordinate system, the transition PDF (e.g. for the x-axis) is

$$\mathrm{f}_{+}(\mathrm{s}_{x+}|\mathrm{s}_x,\ell) = \begin{bmatrix} 1 & t_\Delta \\ 0 & e^{-\beta_x \cdot t_\Delta} \end{bmatrix} \cdot \mathrm{s}_x + w_x \cdot \begin{bmatrix} 0 \\ \sigma_x\sqrt{1-e^{-2\beta_x \cdot t_\Delta}} \end{bmatrix}, \tag{4.1}$$

where $\mathrm{s}_x = [\wp_x, \dot{\wp}_x]^T$, $\dot{\wp}_x$ is the speaker velocity on the x-axis, w_x follows the normal distribution to accommodate the random speaker acceleration and the modelling uncertainty, i.e. $w_x \sim \mathcal{N}(0,1)$ and t_Δ is the time step. Model parameters β_x and σ_x are respectively the rate constant and the steady-state root-mean-square velocity for the random motions of speakers.

4.3.2 Likelihood Function for Multi-sensor Measurements

In the studied scenario, the observation space and the state space are different and a linear transform does not exist between these two spaces. The direct measurements are sets of DOAs, while the states are characterized by the multi-speaker Cartesian locations and velocities, i.e. $\{\mathrm{s}|\mathrm{s} = [\mathrm{s}_x^T, \mathrm{s}_y^T]^T\}$, cf. (4.1). In such case, there are in general two ways to proceed with.

NONLINEAR TRANSFORM

The first method directly uses the measurement space to update the posterior probability densities, if the object birth PDF in (2.104a) is known *a priori*. In this case, the Bayes update can be carried out for measurements from every single microphone array in an "iterated" way. The likelihood function for each microphone array is complicated,

$$g_d^{(a)}(\hat{\theta}|\check{\mathrm{s}}) \sim \mathcal{N}\left(\hat{\theta};\ \arctan\left(\frac{[0,0,1,0]\cdot\check{\mathrm{s}}^{(a)} - \overrightarrow{M}_y^{(a)}}{[1,0,0,0]\cdot\check{\mathrm{s}}^{(a)} - \overrightarrow{M}_x^{(a)}}\right), \sigma_{\hat{\theta}}^2\right),$$

where $\overrightarrow{M}^{(a)} = [\overrightarrow{M}_x^{(a)}, \overrightarrow{M}_y^{(a)}]^T$ is the location of the microphone array a, and $\check{\mathrm{s}}^{(a)}$ is an updated state from the previous “sensor” (i.e. microphone array).

This "iterated" method suffices for a small number of microphone arrays. When the number of microphone arrays is large however (e.g. using multiple TDOA estimates from distributed microphone arrays [4]), the multi-sensor GLMB filter implementation [119, 143] can be applied, to avoid loss of estimates due to the truncation process in each sensor update and hence the dependence on the sequence of the sensor updates.

LINEAR TRANSFORM

The second approach moves the nonlinear transform to a pre-processing step, by converting the direct measurements to a new observation space so that a linear transform exists between the new observation space and the state space. Triangulation is applied to pre-convert at each time instant all the DOA measurements to Cartesian location candidates, which are then filtered by the GLMB recursion. The measurement likelihood function is thus simple,

$$g_c(\hat{z}_{i,j}|\mathrm{s}) \sim \mathcal{N}\left(\hat{z}_{i,j};\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \mathrm{s}, P_{cart}\right), \tag{4.2}$$

where the DOA estimates can be pre-converted to Cartesian locations via triangulation, i.e.

$$\arg(\hat{z}_{i,j} - \overrightarrow{M}^{(a)}) = \hat{\theta}_i^{(a)} \tag{4.3a}$$

$$\arg(\hat{z}_{i,j} - \overrightarrow{M}^{(b)}) = \hat{\theta}_j^{(b)}, \tag{4.3b}$$

where $\overrightarrow{M}^{(a)}$ and $\overrightarrow{M}^{(b)}$ are the locations of microphone arrays, $\hat{\theta}_i^{(a)}$ and $\hat{\theta}_j^{(b)}$ are DOA estimates respectively from the two microphone arrays, $\hat{z}_{i,j}$ the corresponding Cartesian location candidate, and P_{cart} the variance of location in the Cartesian coordinate.

Note that the “nonlinear transform” method still requires a “pre-conversion” step to convert DOAs to Cartesian locations for the initialization of the recursion. Moreover, when using the MDB for adaptive filtering, the birth PDF is drawn from the measurements at every time instant. Thus the measurement-

to-state pre-conversion is actually required for each recursion, which also negates the benefit of using direct DOA measurements for the Bayes update. Moreover, the "iterated" multi-sensor implementation requires a number of sensor iterations in each Bayesian recursion. Therefore, in this thesis, the "nonlinear transform" method is not further pursued, and the "linear transform" is used to pre-convert the DOA estimates to the Cartesian location candidates.

4.4 Numerical Studies

Generally speaking, the performance of the tracking filter largely depends on the knowledge of the state transition function and the measurement accuracy. When the state transition function is completely known, the MDB GLMB filter is capable of adaptively tracking multiple objects in presence of considerable measurement noise and clutter. A detailed example for the multi-object range and bearing tracking can be found in [25]. For multi-speaker tracking however, it would be rather impractical to assume a complete knowledge of the speaker state transition function. Thus the Langevin model (4.1) is often applied to accommodate the practically unpredictable speaker motion. The evaluation of the resulting multi-speaker tracking implementation is provided as follows.

4.4.1 Test Setup

The test set-up for speaker tracking is basically the same as that of the real-world scenario in Section 3.7.1, except now two microphone arrays are used for tracking in Cartesian coordinates.[1] The evaluation uses real audio data recorded in a reverberant office room with measured reverberation time of $T_{60} \approx 0.65$s. The raw speech signals and an actual recording in the room are given in Fig. 3.7. The dimensions of the room is 3.4m$\times$7.6m$\times$2.7m

[1] It has been found in the study that for moving sources the simulation using the method in [62, 140] and overlap-add may create unexpected spurious sources for reverberant conditions. Thus in the thesis only the real-world recordings are used for moving sources.

(width $\times$ length $\times$ height). Two UCAs are used with the radius of $r_a = 0.05$m. Each microphone array has $I_M = 8$ microphones. Microphone arrays are placed close to the centre of the room: [1.2, 3.9, 1.5]m and [2.2, 3.9, 1.5]m, respectively. Fig. 4.2 depicts the room dimensions, locations of two microphone arrays and three speaker trajectories. Three speakers talk and

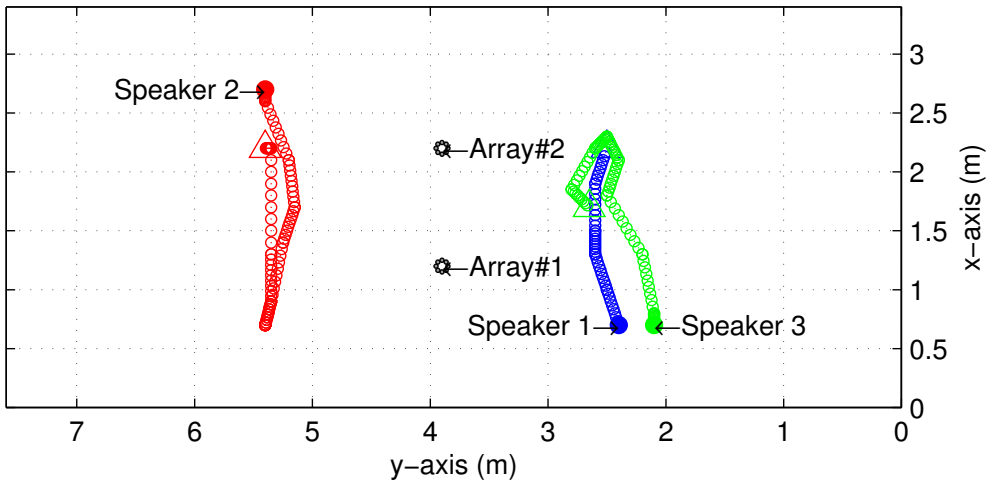

Figure 4.2: Room dimensions (2-D), locations of microphone arrays in black circles, and tracks of speakers in red, blue and green. Starting locations are solid circles and ending locations are triangles.

move in the room, and the number of active speakers at any time instant is unknown to the algorithm and changes over time. One of the speakers is female while the other two are male. The trajectories of speakers are nonlinear.

Parameters are chosen as follows. $\lambda_B = 0.3$, $r_{B_{\max}} = 0.15$, and $M_b = 30000$ for the MDB model; $p_S = 0.98$, $p_D = 0.9$, $\beta_x = 0.5/s$, $\sigma_s = 0.5m/s$, $\kappa = 0.08$ uniformly distributed over the room. From Fig. 3.19, choosing the DOA estimation error as 8° for moving speakers, the location deviation at a distance of 1m is about $(1\text{m} \cdot 8^\circ \pi/180^\circ) \approx 0.15$m. Thus the variances are chosen as $P_B = 0.15^2 \cdot \text{diag}(1m, 1m/s, 1m, 1m/s)^2$ in (2.112), and $P_{cart} = (0.15m)^2 \cdot \text{diag}(1,1)$ in (4.2) for the MDB GLMB filter. The time step used is $t_\Delta = 0.5$s, which is sufficient for common speaker tracking. For performance comparison, the SMC-CPHD [19] and its adaptive birth model [120] is also

implemented with common parameters of the same values with the GLMB.

4.4.2 Evaluation Results For Speaker Tracking

DOA Estimates

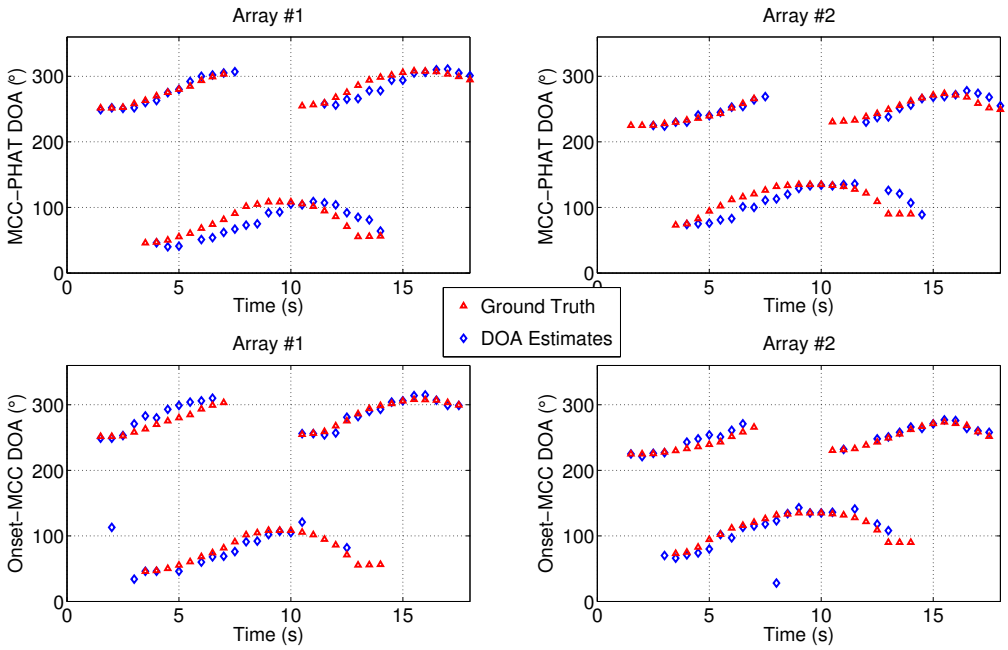

Figure 4.3: Ground truth and estimated speaker DOAs from two microphone arrays ($T_{60} \approx 0.65s$), using MCC-PHAT and Onset-MCC methods.

The DOA estimates from the proposed localization methods are plotted in Fig. 4.3, in comparison with the ground truth. As shown in the plots, there are spurious peaks, miss-detections and offsets due to reverberation, noise and estimation process. However, most of the DOAs are captured by the proposed localization algorithms. The results from the MCC-PHAT and Onset-MCC methods are close in general, except that the MCC-PHAT produces less miss-detections but more offset, while the Onset-MCC has more clutter and miss-detections, but less offset. The miss-detections may result in gaps in the location estimates as can be seen at around 12s. Heuristically the locations can be connected over time to form separated tracks (using e.g. nearest neighbour methods), but this thesis focuses on the applications of

the state-of-the-art GLMB filter, which is a *closed-form* multi-object tracking filter. The SMC implementations of respective multi-object filters are used.

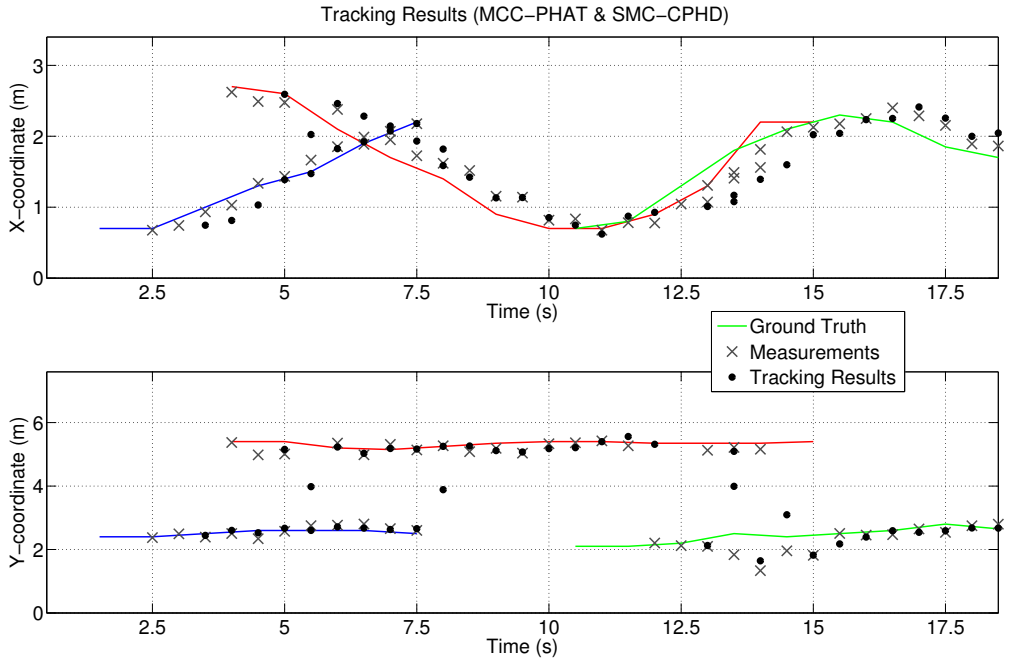

Figure 4.4: Estimated Cartesian tracks of speakers using MCC-PHAT DOA measurements (pre-converted) via SMC-CPHD.

TRACKING RESULTS

Fig. 4.4 and Fig. 4.5 demonstrate the estimated filtering results of speakers from the MDB SMC-CPHD using the triangulation results (4.3) of the DOA estimates from the MCC-PHAT and the Onset-MCC methods, respectively. Same parameters are used in both cases for the MDB SMC-CPHD filter. Obviously, the CPHD does not associate the filtered locations with respective speakers, neither does it establish connections of the filtered locations over time.

Fig. 4.6 and Fig. 4.7 demonstrate the estimated trajectories of speakers from the MDB SMC-GLMB using the triangulation results (4.3) of DOA localization results from the MCC-PHAT and the Onset-MCC methods, respectively. Same parameters are used in both cases for the MDB SMC-GLMB

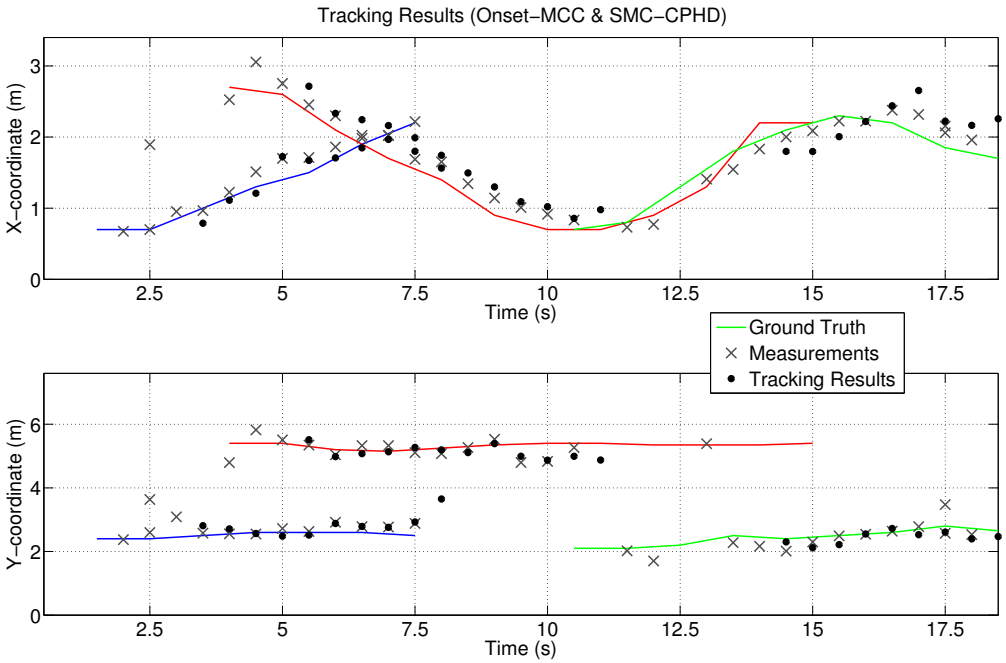

Figure 4.5: Estimated Cartesian tracks of speakers using Onset-MCC DOA measurements (pre-converted) via SMC-CPHD.

filter. Both figures show that the tracking results are close to ground truth. Compared with CPHD results, not only the location estimates of speakers are provided, but also the trajectories are assigned with separate labels (identities) that correspond to different speakers. Thus in what follows, focus is kept on the evaluations of the GLMB implementations.

In Fig. 4.6, the short gaps (e.g. Speaker 2 at around 12s) did not break the trajectories, as the location candidates before and after the gaps still follow the kinematic models. In Fig. 4.7, the long gaps (Speaker 3 at around 12s) from the Onset-MCC however, result in lost of part of the track. This lost of track can be fixed by some heuristic track management step, but is not in the scope of the thesis. Thus the MCC-PHAT seems preferable to the Onset-MCC in extracting speaker trajectories. It can also be seen that it takes about two time steps to confirm a new track, mainly because that the MDB models implemented assume no *a priori* knowledge of objects births. The computational complexity analysis of the tracking methods is provided in Appendix K .

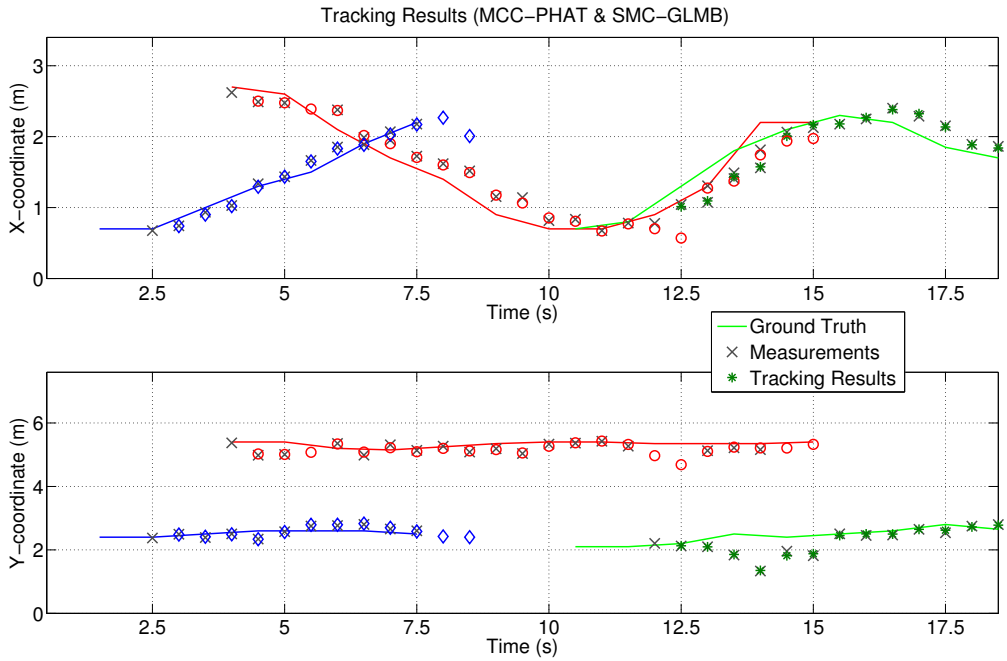

Figure 4.6: Estimated Cartesian tracks of speakers using MCC-PHAT DOA measurements (pre-converted). Tracks with different labels are plotted with different colors and symbols.

OSPA Performance

OSPA results of the GLMB tracking are given in Fig. 4.8 and Fig. 4.9, respectively. Parameters $p = 2$ and $c = 0.6m$ are chosen here for evaluations. The average Cartesian errors of the proposed algorithm are around $0.5m$, which is lower than the common inter-person distance. The speakers are correctly identified with different labels. The beginning and ending time and durations of speaker activities are close to the ground truth. The OSPA metric (cf. Section 2.6.6) provides three measures, viz. the overall OSPA distance, the OSPA location error, and the OSPA cardinality error. From Fig. 4.8 that at time 2s, a cardinality error of 1 out of 1 converts to $c = 0.6$m, while a cardinality error of 1 out of 2 at time 7.5s converts to $(\frac{1}{2}(2-1)c^p)^{\frac{1}{p}} = 0.4$m. The overall OSPA distances include both the location offsets and the cardinality errors as defined in (2.113).

The Monte Carlo test is carried out to verify the consistency by repeating

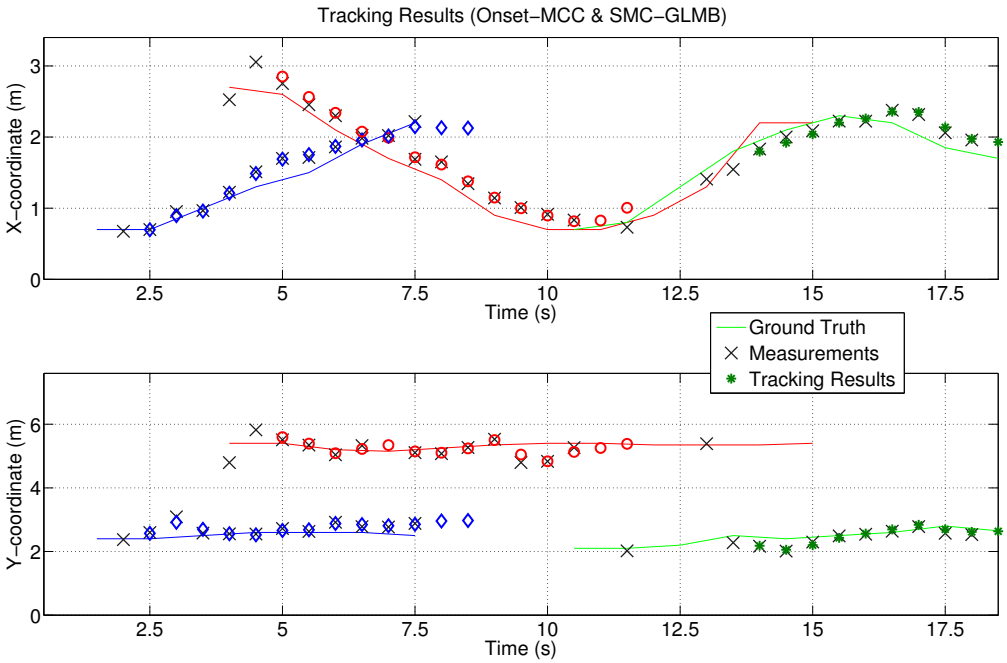

Figure 4.7: Estimated Cartesian tracks of speakers using Onset-MCC DOA measurements (pre-converted). Tracks with different labels are plotted with different colors and symbols.

the tracking algorithm for 500 times. The maximal deviations are also plotted. In both cases, most of the time the maximum and minimum OSPA errors overlap respectively. Thus the proposed system works consistently.

4.5 Conclusions

This chapter presents a new framework for adaptively tracking an unknown and time-varying number of moving speakers in highly reverberant environments with nonstationary speech signals. The proposed framework consists of two stages, namely the acoustic feature extraction (i.e. localization, cf. Part I), and the Bayes RFS multi-object tracking filter (GLMB v.s. CPHD). The acoustic feature extraction is based on the proposed MCC-PHAT and Onset-MCC reverberation-robust localization methods from Part I. The Bayes RFS tracking is implemented using the GLMB filter with the MDB model, and supplied with multi-sensor measurements.

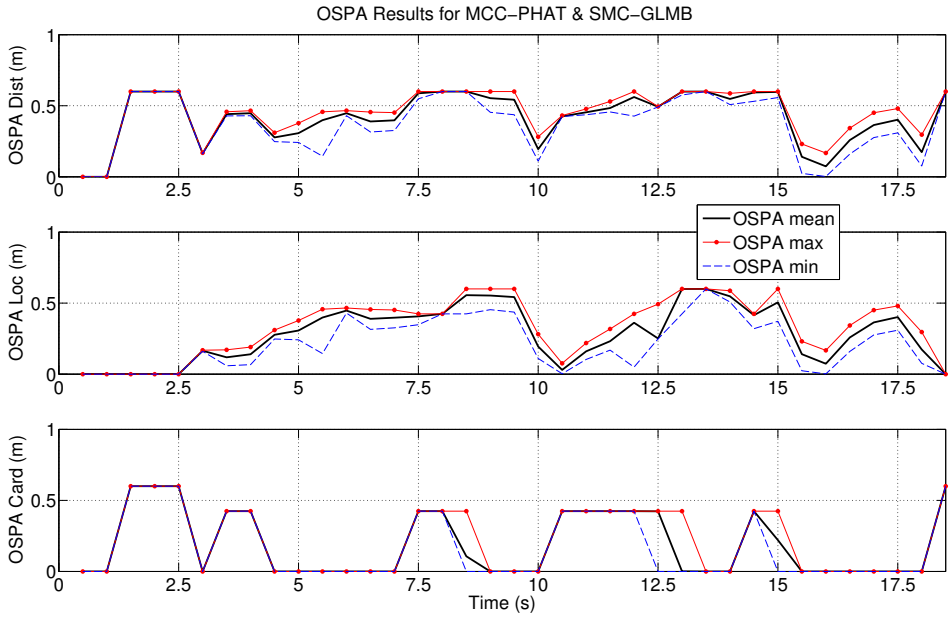

Figure 4.8: OSPA Results using MCC-PHAT and SMC-GLMB.

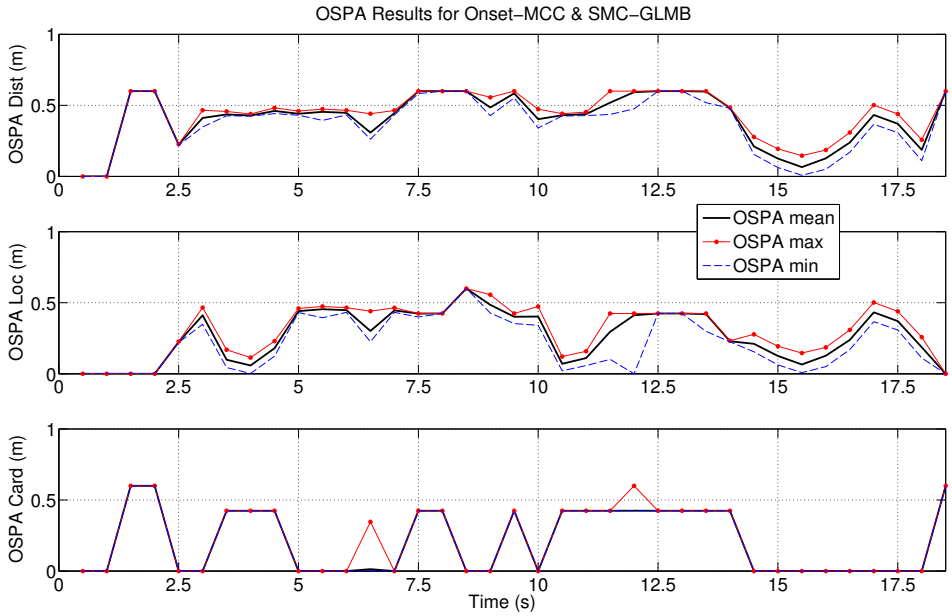

Figure 4.9: OSPA Results using Onset-MCC and SMC-GLMB.

Performance of the proposed multi-speaker tracking framework is demonstrated using real recordings in a reverberant room at $T_{60} \approx 0.65$s, with the

scenario where one female speaker and two male speakers talk and move in a reverberant room. As the results show, the labeled trajectories are reasonably close to ground truth, despite of the significant challenges, such as reverberation, moving speakers, time-varying number of speakers, and joint labeled state tracking. The MCC-PHAT and Onset-MCC can both reliably find speaker DOAs in presence of strong reverberation, and the MCC-PHAT has comparatively less miss-detections. The MDB SMC-GLMB filter, supplied with location measurements, estimates the kinematic states of multiple moving speakers jointly with identities. Comparison with the CPHD counterparts also confirms the advantages of the GLMB filter implementations.

Chapter 5

Adaptive Multi-feature Multi-speaker Tracking and Separation

As presented in Chapter 4, standard multi-speaker tracking algorithms usually only filter the kinematic feature state. In such implementations, ambiguity arises when speakers are spatially close, which cannot be resolved without using other features. Thus this chapter investigates the feasibility of jointly tracking and separating multi-feature states of multiple speakers, based on the MDB GLMB multi-object tracking filter. A *multi-feature* multi-speaker *tracking-and-separation* method is proposed, using sound mixtures recorded by microphones. The proposed multi-feature GLMB tracking filter treats the set of vectors of speaker features (e.g. location, pitch and sound) as the multi-feature multi-object observations, characterizes transitioning features with corresponding transition models and overall measurement likelihood function, thus jointly tracks and separates each multi-feature speaker, and addresses the ambiguity problem of single feature state filters. As a proof-of-concept, this chapter uses a simulated anechoic scenario to verify that the proposed method can correctly track locations of multiple speakers and meanwhile separate speech signals. Main contributions of this chapter have been published in the author's paper[1] [6].

5.1 Introduction

Numerous acoustic multi-speaker tracking algorithms can be found in the literature [2, 4, 142, 144]. Generic multi-object tracking filters [19, 21–23] can also be implemented to track multiple speakers online when provided with speaker location estimates as observation data. These existing implementa-

[1] ©2018 IEEE. Reprinted with permission from, Jointly Tracking and Separating Speech Sources Using Multiple Features and the generalized labeled multi-Bernoulli Framework, by Shoufeng Lin. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018).

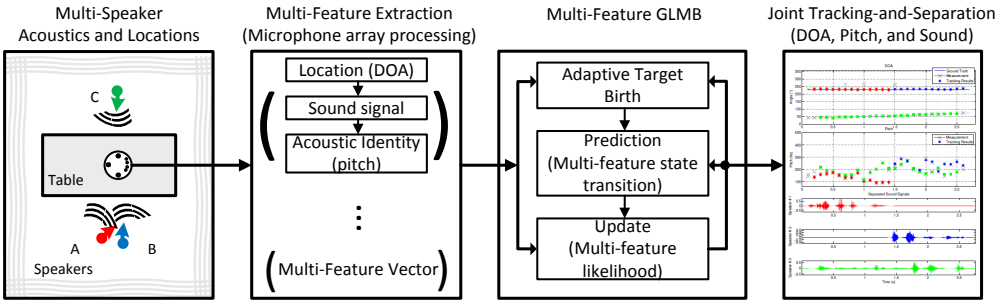

Figure 5.1: Multi-feature Multi-speaker Tracking and Separation. ©2018 IEEE.

tions of multi-speaker tracking methods however, usually track only a single (kinematic) state of respective speakers. Nonetheless, single feature tracking has the ambiguity problem when the feature observations are close to each other. Apparently, by relying on the location information alone, the tracking filters would take closely-located speakers as a single speaker, hence unable to correctly identify and separate the sound sources in the mixture.

Separating original source signals from the mixtures recorded by microphones has also a wide range of applications such as automatic meeting transcription and speaker recognition. Many blind source separation (BSS) methods have been developed [50, 145–147], based on the independent component analysis (ICA) or time-frequency masking (TFM) techniques. However, it can be challenging for some BSS methods to continuously separate moving sources due to the well-known permutation problem. Thus the location-based source separation methods, e.g. the wideband beamforming methods (e.g. [10, 31, 76]), are often employed as an additional source separation step after obtaining the location tracking results.

This chapter proposes a systematic multi-feature tracking-and-separation filter based on the generalized labeled multi-Bernoulli (GLMB) framework [21–23]. As shown in Fig. 5.1 (and cf., Fig. 4.1), multiple speaker features are first obtained from sound mixtures by detecting locations of all candidate speakers, extracting their corresponding speech signals and estimating the related acoustic identities (pitches). Each extracted vector of associated speaker features of a candidate speaker, i.e. the location, pitch and the cor-

responding speech signals, can be treated as an integral multi-feature target observation. The set of multi-feature vectors forms the multi-target multi-feature observations, which are then tracked in the proposed multi-feature GLMB. Moreover, since the standard implementations of the GLMB framework [21–23] track only a single feature, necessary adaptations are required to support multi-feature tracking. The proposed method categorizes the location and pitch as "transitioning" features, and the nonstationary sound signal as a "non-transitioning" feature. In the multi-feature GLMB recursion, transitioning features have their own first-order Markov transition models and are directly used for track confirmation in the update step, while the non-transitioning sound feature is zeroed (as silence) in the prediction step and assigned with associated extracted sound in the update step. New state transition function and measurement likelihood function for multiple transitioning features are also presented. The multi-feature GLMB tracking filter produces labeled tracks for respective speakers, the corresponding pitch estimates, as well as the separated sound signals. Furthermore, it also addresses the ambiguity problem because when speakers locate closely, their pitch information can be used to separate them in the multi-feature GLMB tracking algorithm, and vice versa.

This chapter is organized as follows. Section 5.2 describes the speaker feature extraction methods. The multi-feature GLMB filter is described in Section 5.3. Numerical studies of the proposed methods are presented in Section 5.4, and conclusions are given in Section 5.5.

5.2 Speaker Feature Extraction

5.2.1 Speaker Localization

All localization methods as discussed in Section 2.5 or proposed in Chapter 3 can be applied at this stage. Here the MCC-PHAT method as described in Section 3.5.4 is applied for a satisfactory overall performance. The location estimates are as denoted in (3.62), i.e. $\hat{\Theta}_k = \{\hat{\theta}_{i_k} \mid i_k = 1, \ldots, N_k\}$.

5.2.2 Sound Extraction

Speech signals from the DOA estimates $\hat{\theta}_{i_k}$ can then be extracted from the sound mixtures recorded by microphones. Here the wideband weighted least square (WLS) beamforming method [31] as described in Section 2.5 is implemented for sound extraction.

The WLS beamformer uses the filter-and-sum structure, and has $J = 32$ taps in each channel. Its mainlobe steers to the speaker DOA $\hat{\theta}_{i_k}$, and the corresponding sidelobe ranges from $\hat{\theta}_{i_k} + 15^\circ$ to $\hat{\theta}_{i_k} - 15^\circ$. The frequency range used is [20, 8000]Hz.

The real-valued $(J \cdot I_M) \times 1$ optimal weight vector $\mathbf{w}_{i_k}$ for a DOA $\hat{\theta}_{i_k}$ is obtained according to the wideband WLS beamformer [31], then the extracted sound signal at time frame k can be calculated from (cf. (2.49)):

$$\hat{y}_{i_k}[n] = \mathbf{w}_{i_k}^T \mathbf{x}_{BF}[n]. \tag{5.1}$$

5.2.3 Acoustic Identity

The extracted sound $\hat{y}_{i_k}$ that corresponds to a speaker location $\hat{\theta}_{i_k}$ can further be used to extract speaker's acoustic identity, e.g. pitch, Gaussian Mixture Model (GMM) [148] parameters, etc. Here the pitch is used as a simple acoustic identity, as it can be estimated from a short segment of voiced sound, different speakers usually have different pitch, and pitch of a speaker is usually distributed within a limited range. Numerous pitch estimation methods can be found in the literature [106, 149–152]. The PEFAC (Pitch Estimation Filter with Amplitude Compression) method [151] is employed. The averaged estimate of each frame is used, which is denoted as $\hat{F}_{0i_k}$.

From (3.62) and (5.1), the vector of associated location, pitch and sound of each candidate speaker at frame k form a multi-feature observation $z_{i_k} \triangleq (\hat{\theta}_{i_k}, \hat{F}_{0i_k}, \hat{y}_{i_k})$. The multi-target multi-feature observation is thus

$$Z_k \triangleq \{z_{i_k} \mid i = 1, ..., N_k\}, \tag{5.2}$$

where $Z_k = \emptyset$ when $N_k = 0$.

Instead of using the location estimates alone, the proposed multi-feature GLMB filter jointly extract and track the location, pitch and sound features as follows.

5.3 Multi-feature GLMB Recursion

The multi-feature GLMB RFS is the same form as (2.99), except that here $\mathrm{s}_i \triangleq (\theta_i, F_{0i}, y_i) \in \mathbb{X}$ is the multi-feature target state vector, where θ_i, F_{0i}, y_i denote the associated location and pitch feature states as well as the sound waveform, respectively. The multi-feature GLMB recursion also consists of the multi-object "update" step based on Bayes inference and the Chapman-Kolmogorov [117] "prediction" step based on the state transition models.

5.3.1 Update

The form of multi-feature GLMB update is the same as (2.101). However, the multi-feature likelihood function should be adapted.

$\mathrm{g}(z_{\vartheta(\ell)}|\mathrm{s},\ell)$ denotes the multi-feature likelihood for the measurement $z_{\vartheta(\ell)} \in Z$ being generated by $(\mathrm{s},\ell) = ((\theta, F_0, y),\ell)$, where the feature y is "non-transitioning" and assigned with $\hat{y}_{\vartheta(\ell)}$ after update. Sound separation for respective speakers over time is achieved by concatenating sound signals y of the same target label. Assuming that the transitioning features (location and pitch) are statistically independent, the proposed multi-feature likelihood function is:

$$\mathrm{g}(z_{\theta(\ell)}|\mathrm{s},\ell) \triangleq \mathrm{g}(\hat{\theta}_{\vartheta(\ell)}|\theta,\ell) \cdot \mathrm{g}(\hat{F}_{0\vartheta(\ell)}|F_0,\ell), \tag{5.3}$$

where $\mathrm{g}(\hat{\theta}_{\vartheta(\ell)}|\theta,\ell) = \mathcal{N}(\hat{\theta}_{\vartheta(\ell)};\theta,\sigma_\theta^2)$ and $\mathrm{g}(\hat{F}_{0\vartheta(\ell)}|F_0,\ell) = \mathcal{N}(\hat{F}_{0\vartheta(\ell)};F_0,\sigma_{F_0}^2)$. $\sigma_\theta = 2^\circ$ and $\sigma_{F_0} = 10$Hz are the standard deviations of the observation of the location and pitch, respectively. Following the definitions in [22], clutter is assumed Poisson with an average of 0.044 clutter points per scan, i.e. the localization method produces almost clean location estimates in low reverberation. The probability of a target state being detected is $p_D = \mathcal{N}(F_0;280,30^2)$.

After update, the maximum *a posteriori* (MAP) estimate of the cardinality (number of speakers) is chosen, and the highest weighted corresponding hypothesis is used for the multi-target multi-feature tracking results.

5.3.2 Prediction

The form of multi-feature GLMB prediction is the same as (2.103). However, the multi-feature transition function must be adapted.

Using the assumption that transitioning features are statistically independent, the proposed state transition function for the multi-feature GLMB is

$$\mathrm{f}(\mathrm{s}|\cdot,\ell) = 1_{\mathrm{s}}(\theta)\cdot \mathrm{f}(\vec{\theta}|\cdot,\ell) \;\cdot\; 1_{\mathrm{s}}(F_0)\cdot \mathrm{f}(F_0|\cdot,\ell), \tag{5.4}$$

where the inclusion function is as defined in (2.105). The survival probability here is $p_S(\cdot,\ell) = 0.75$, considering the dynamics of multiple features.

Assume that the speaker DOA follows the Langevin process with the same form defined in (4.1) but with different parameter values,

$$\mathrm{f}(\vec{\theta}|\vec{\theta}',\ell) = \begin{bmatrix} 1 & t_\Delta \\ 0 & e^{-\beta_\theta\cdot t_\Delta} \end{bmatrix}\cdot\vec{\theta}' + w_\theta\cdot\begin{bmatrix} 0 \\ \sigma_\theta\sqrt{1-e^{-2\beta_\theta\cdot t_\Delta}} \end{bmatrix}, \tag{5.5}$$

where $\vec{\theta} = [\theta,\dot{\theta}]^T$, $\dot{\theta}$ is the velocity of DOA θ. $t_\Delta = 0.1$s is the time step, w_θ follows the normal distribution, i.e. $w_\theta \sim \mathcal{N}(\cdot;0,1)$. Model parameters $\beta_\theta = 0.2\mathrm{s}^{-1}$ and $\sigma_\theta = 10^\circ/\mathrm{s}$ are respectively the rate constant and the steady-state root-mean-square velocity for the random motions of speakers.

Assume that the pitch of a speaker follows a simple normal distribution around its previous estimate. Thus the state transition function for pitch is

$$\mathrm{f}(F_0|\mathrm{F_0}',\ell) = \mathcal{N}(F_0;F_0',\tilde{\sigma}_{F_0}^2), \tag{5.6}$$

where $\tilde{\sigma}_{F_0} = 30$Hz is the standard deviation for the transition of pitch. Adaptive measurement-driven target births are generated [23, 25]. New target births are assumed to follow normal distributions around the previous measurement, where the standard deviation is 5° for the DOA (cf. Part I, Sec-

tion 3.7.2), and 30Hz for the pitch, respectively. The nonstationary sound signals are treated as the non-transitioning feature, thus targets carry no sound in prediction until the next update step of the multi-feature GLMB recursion.

5.4 Numerical Studies

5.4.1 Experiment Setup

This section verifies and demonstrates the performance of the proposed multi-feature GLMB framework in the scenario of three speakers.

The test set-up is as shown in the left panel of Fig. 5.1, where the room dimensions are $3.4m \times 7.6m \times 2.7m$ (width $\times$ length $\times$ height), the microphone array locates at [1.2, 3.9, 1.5]m, which is composed of $I_M = 8$ microphones evenly distributed on a circle with a diameter of 0.1m. For clarity, an anechoic scenario is chosen that Speaker A (male) and B (female) both locate at DOA of 232.1° while Speaker C (female) moves from DOA of 40° to 75°, with respect to the center of the microphone array. Fig. 5.2 plots the normalized ground truth speech signals of respective speakers as well as their mixture captured by one of the microphones. Obviously, using location (DOA) information alone, standard implementations of tracking methods can only take Speaker A and B as a same speaker.

5.4.2 Test Results

Fig. 5.3 provides the ground truth locations, estimated speaker locations, pitch and separated sound signals. The top panel depicts the ground truth locations in straight line segments, the estimated locations in symbol "$\times$" and tracking results in solid colored symbols. Different colored symbols represent different speakers. From the ground truth, there are two separate lines of locations. Thus using location information alone, apparently the tracking filters can only detect two speakers. However, by considering also the pitch information, the proposed method has correctly found three speakers. The

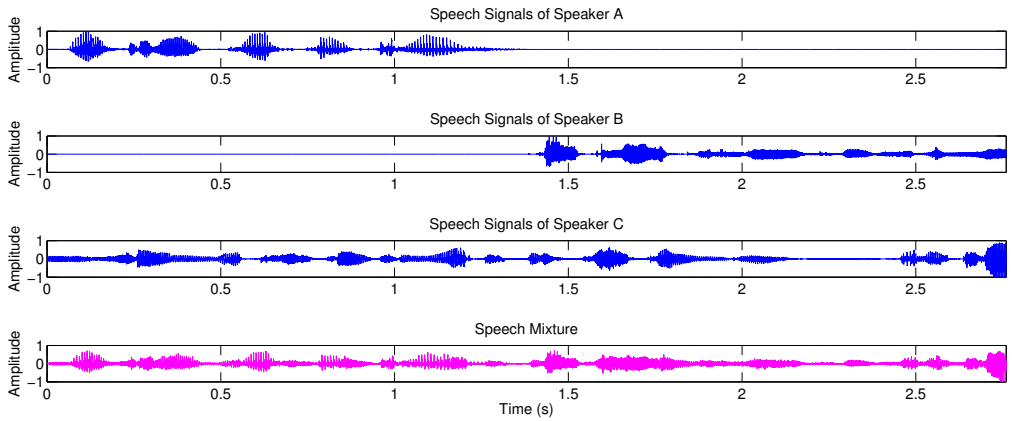

Figure 5.2: Ground truth (top three panels) of the normalized speech signals of three speakers (one male and two female), and their mixture at one of the microphones (bottom panel). ©2018 IEEE.

second top panel shows the pitch estimates and tracking results associated with the location estimates and tracking results in the top panel. In these two panels, the associated location and pitch estimates have spurious errors that do not follow consistent kinematic patterns over time, thus are filtered by the GLMB tracker. The tracking filter requires two time steps to confirm one new track. This is reasonable as the measurement-driven birth model [25] is used for adaptive target births. The pitch estimates of different speakers fluctuate at different levels over time, and there is a significant jump in pitch level at time of around 1.4s, which helps the tracker to confirm a new speaker starting at 1.5s. The bottom three panels of Fig. 5.3 plots the extracted sound signals for respective speakers. Comparing with Fig. 5.2, most of speech signals are recovered for each speaker. Thus the proposed multi-feature GLMB tracking-and-separation method can jointly track and separate multiple speakers.

The location tracking accuracy is evaluated using the OSPA metric [122], with the cut-off parameter of 5° and the order parameter of 1. Thus cardinality estimation error of 1 out of 2 contributes to an OSPA error of $\frac{5}{2}^\circ$. Fig. 5.4 shows that the overall OSPA location tracking errors are within 5°, and the multi-feature GLMB achieves comparable location tracking accuracy with the standard GLMB.

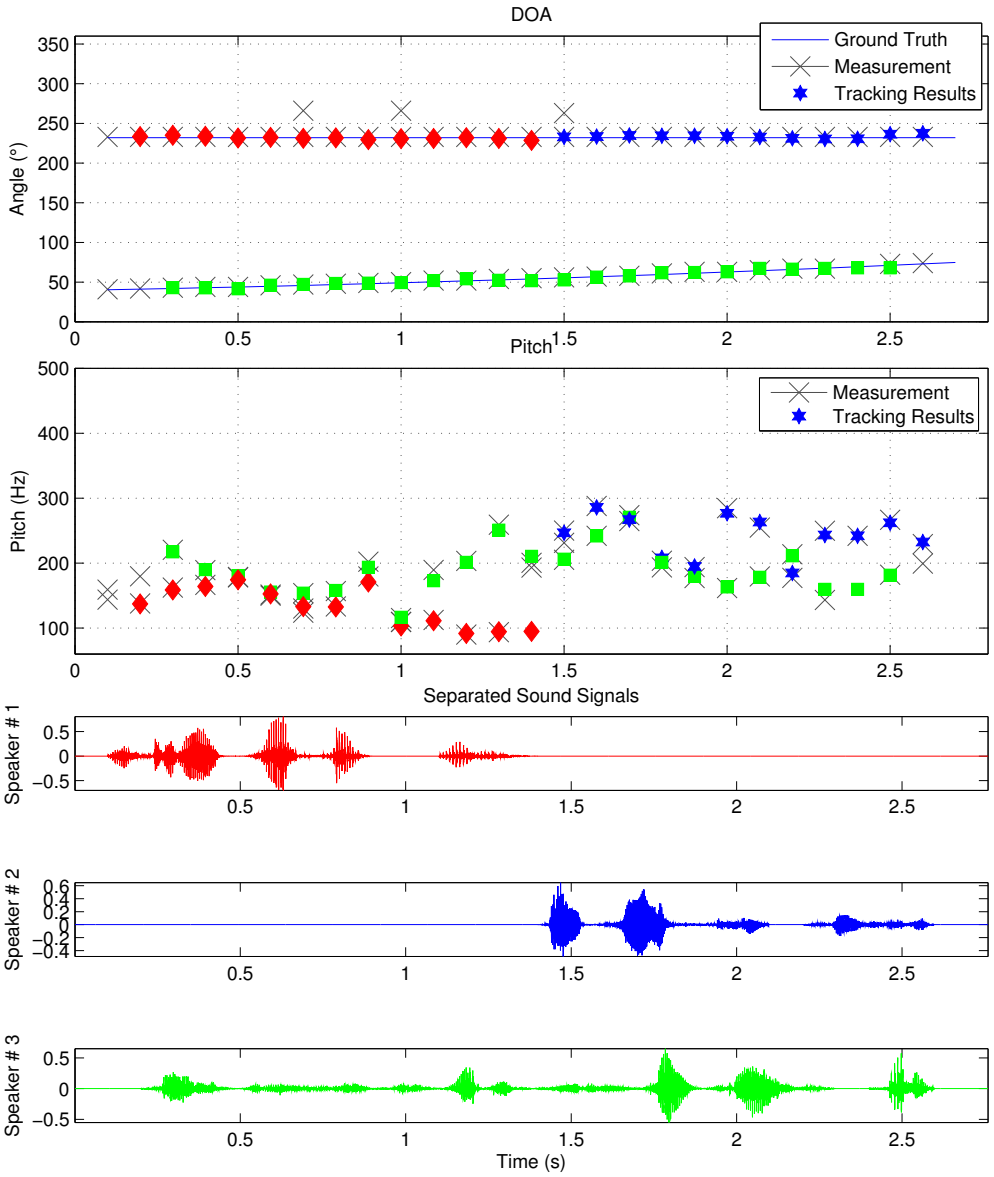

Figure 5.3: Joint tracking and separation results from proposed methods. Top two panels show the estimation and tracking results of speakers' location and pitch. Bottom three panels show the corresponding separated sound signals. ©2018 IEEE.

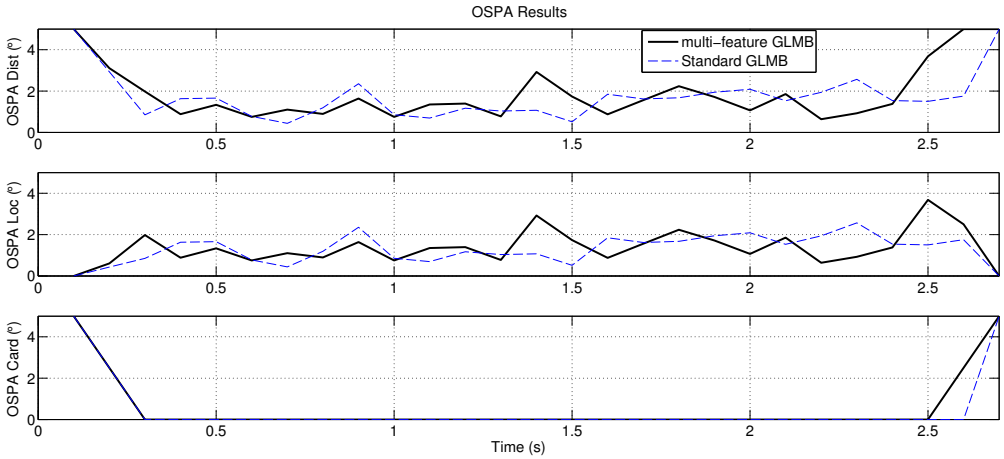

Figure 5.4: OSPA measure of the DOA tracking results, i.e. the overall OSPA errors (top), the contribution of DOA errors (middle), and the contribution of cardinality errors (bottom). ©2018 IEEE.

Table 5.1: PEASS evaluation results for speech separation, using the proposed method, and the UCBSS, DUET methods. ©2018 IEEE.

Method	Speaker	OPS	TPS	IPS	APS
Proposed	1	48.75	57.03	71.19	49.11
	2	32.69	29.35	72.06	35.61
	3	36.02	35.73	65.65	37.71
UCBSS	$<1,2>$	18.66	45.84	43.21	24.33
	3	25.00	6.10	83.97	3.50
DUET	$<1,2>$	18.73	38.82	16.38	50.43
	3	24.97	51.16	32.40	44.32

The quality of the separated sound signals are evaluated using the PEASS metric [153], compared with the ground truth signals. The results are provided in Tab. 5.1. The performance is compared with two blind speech separation methods, i.e. the Underdetermined Convolutive Blind Source Separation (UCBSS) [147] and the Degenerative Unmixing Estimation Technique (DUET) [50]. Using the blind separation techniques, the speaker 1 and speaker 2 are regarded as one speaker. Thus the separated sound signals for speaker $<1,2>$ are compared with the mixture of Speaker A and Speaker B. In general the DUET and UCBSS methods obtain close Overall Perceptual Scores (OPS). The DUET method seems to provide more con-

sistent performance than UCBSS when comparing the Target-related Perceptual Score (TPS) and the Artifacts-related Perceptual Scores (APS), but UCBSS has significantly higher Interference-related Perceptual Score (IPS) than DUET. Overall, the proposed method provides consistent and superior performance for the three separated speakers, according to the perceptual scores.

5.5 Conclusions

This chapter presents a novel systematic implementation of the multi-feature multi-speaker filtering method that not only can jointly track multiple speakers and separate sound signals from speech mixtures, but also can resolve the ambiguity of location tracking when speakers locate spatially close. It treats the vector of candidate speaker location, pitch and sound as a multi-feature target observation and jointly extracts and tracks these features in the GLMB Bayes RFS recursion. Experimental results demonstrate encouraging results in the studied scenario.

Part III

Conclusions and Future Works

Chapter 6

Conclusions and Future Works

6.1 Conclusions

The thesis investigates the challenges in the reverberant speaker localization and multi-speaker feature filtering and presents several original approaches.

Three novel reverberation-robust speaker localization algorithms are presented, which are referred to as the Onset-GSRP, Onset-MCC and the MCC-PHAT. The Onset-GSRP and Onset-MCC algorithms first decompose speech mixtures via the auditory filterbank based on the voiced speech signal model and time-frequency sparsity assumption. Then a novel onset detection and encoding approach is derived to detect the direct-path components from reverberant microphone recordings, based on the voiced speech signal and acoustic RIR models. Finally, the cross-correlation coefficients are reverse-mapped from relative sample delays to spatial locations, and produce the DOA localization results in a computationally tractable way. The MCC-PHAT method builds upon the acoustic RTF model and the classic GCC-PHAT method, and exploits the redundant information from multiple closely placed microphones to suppress the effect of sound reflections. Performance of the presented methods is studied using not only simulated signals of reverberation time from 0.2 to 1s, but also real recordings in an office room of $T_{60} \approx 0.65$s. Evaluation results show that the proposed Onset-MCC and MCC-PHAT speaker localization methods can provide better DOA resolutions than the Onset-GSRP and other baseline techniques. Comparative studies demonstrate the benefits of the proposed algorithms, i.e. good resolutions, reverberation-robustness and localization of moving speakers.

Moreover, an adaptive multi-speaker tracking filter is developed for an unknown and time-varying number of moving speakers in highly reverberant environments using microphone recordings of nonstationary speech signals. The proposed filter consists of two stages, namely the acoustic feature (locations) extraction, and the GLMB Bayes RFS multi-object tracking. The

acoustic feature extraction uses the proposed Onset-MCC and MCC-PHAT reverberant-robust localization methods. The GLMB Bayes RFS filter is implemented with the MDB model, and supplied with pre-converted Cartesian location candidates from multi-sensor DOA estimates. Performance of the proposed multi-speaker tracking framework is evaluated using real recordings in a reverberant room at $T_{60} \approx 0.65$s, with the scenario where three speakers talk and move in a reverberant room. The results indicate that the MDB GLMB filter, supplied with location estimates, can adaptively track the kinematic states of multiple moving speakers jointly with identities. Comparison with the CPHD filter also confirms the advantages of the GLMB filter implementations.

Furthermore, a novel multi-feature multi-speaker filter is also proposed, based on the GLMB Bayes RFS filter framework. It treats the vector of candidate speaker location, pitch and sound as a multi-feature observation, and jointly filters (i.e. tracks and separates) these features in the MDB GLMB Bayes RFS recursion. Experimental results demonstrate that the proposed multi-feature multi-speaker filter can jointly track the locations and pitches of multiple speakers and separate corresponding sound signals from speech mixtures in the studied scenario.

6.2 Future Works

1. UCAs can provide isotropic localization performance (see e.g. [87]). The works on theoretical performance bounds and variances of location estimators using the redundant information, especially in reverberant conditions and for moving sources, as well as the posterior performance bounds for state estimation, are not included in the thesis. Existing literature (e.g. [100, 154–158]) also indicated the significance and the level of challenges of a further study in this regard. Moreover, since the thesis focuses on the challenges of reverberation and moving speakers, further theoretical and practical performance studies for static speakers can be carried out next for certain applications.

2. The multi-feature multi-speaker filter has been verified in Chapter 5, but since some estimators of the features (e.g. the pitch estimator and the beamformer for speech separation) are not very reliable in presence of reverberation, the test scenario is in an anechoic condition, and only a proof-of-concept is given. Moreover, the more challenging scenario when closely located speakers talk concurrently is to be investigated with better speech separation methods. In future works, reverberation-robust pitch estimators can be investigated, which may help in the joint tracking and separation in more challenging multi-speaker scenarios, e.g. the "cocktail party" problem [159–161]. Furthermore, the framework provides a basis for following automatic speech recognition (ASR) and natural language processing (NLP) in practical scenarios as well.

3. Fixed UCAs are used in the thesis for testing proposed and existing methods. There are also cases when the sensor arrays are also moving. In such scenarios, further investigation is needed.

4. An integrated multi-sensor multi-object filtering [143] can be implemented for the case of multiple microphone arrays (e.g. distributed array network), although not necessary in this thesis as only two microphone arrays are used, and the source code is not yet released by the authors of the paper.

5. A simple fixed wideband beamformer is used for speech separation in the thesis. Future works may include more adaptive speech separation methods that are robust against reverberation.

6. Simulation of reverberated sound [62, 140] for moving speakers can be improved in future studies. The current simulation found that the overlap-add of simulated reverberated signals can produce unnatural sounds and spurious localization errors. The fundamental cause of this issue may be the conflict of the fixed sampling rate and the continuous motion of speakers. Further investigation can be interesting.

7. The speaker localization and tracking methods proposed in this thesis

can be implemented online. For offline implementations, state smoothing is applicable, which can possibly lead to more accurate results. Other extensions to the tracking framework and transition models are possible for certain complicated scenarios in practice.

8. Microphone arrays may not be perfect, e.g. there may be perturbations in microphone gain, phase, and locations. Thus automatic array calibration and robustness of proposed methods in such conditions also deserves further investigation.

9. Implementing the works in real embedded systems (including hardware and software) will also be of practical values. There are an increasing number of smart home/meeting/industry and wearable devices that play an important part in the daily life. With over 8 years of professional engineering background in electronics products development, the author is also keen on this.

Appendices

A Transfer Function for Room Acoustics

From the law of conservation of momentum (also known as the Euler's equation),

$$\nabla \mathrm{p} + \rho_0 \frac{\partial \mathbf{v}}{\partial t} = 0, \tag{7.1}$$

where ∇ denotes gradient, p the sound pressure, which is the difference between the instantaneous pressure and the static pressure p_0, $\mathbf{v}$ the particle velocity, and ρ_0 the static air density.

From the law of conservation of mass, in free space,

$$\rho_0 \mathrm{div}\mathbf{v} + \frac{\partial \rho}{\partial t} = 0, \tag{7.2}$$

where ρ is the total air density, and div the divergence.

Assuming ideal gas,

$$\frac{\mathrm{p}}{\mathrm{p}_0} = \mathsf{k}_a \frac{\delta \rho}{\rho_0}, \tag{7.3}$$

where k_a is the adiabatic exponent ($\mathsf{k}_a = 1.4$ for air), and $\delta\rho$ the variation of air density. The velocity of sound $v = \sqrt{\mathsf{k}_a \cdot \mathrm{p}_0/\rho_0}$.

From (7.1), (7.2) and (7.3), it is easy to derive the Helmholtz equation

$$\nabla^2 \mathrm{p} = \frac{1}{v^2} \frac{\partial^2 \mathrm{p}}{\partial t^2}. \tag{7.4}$$

For a point source with a single frequency Ω and volume velocity $Q(\Omega) = \hat{Q} e^{\jmath \Omega t}$, the solution to (7.4) is spherical wave, and at a distance of $r_d > 0$ in polar coordinates [61],

$$\mathrm{p}(r_d, t) = \jmath \Omega \rho_0 \hat{Q} \frac{e^{\jmath(\Omega t - k_\lambda r_d)}}{4\pi r_d}, \tag{7.5}$$

where $k_\lambda = \Omega/v$ is the wave number. This leads to the transfer function (2.9).

In an enclosed space with volume V, assuming the single frequency point

source with a vanishingly small volume dV locate at location $\vec{\wp}_0$, (7.2) is rewritten as

$$\rho_0 \mathrm{div}\mathbf{v} + \frac{\partial \rho}{\partial t} = \rho_0 q(\Omega, \vec{\wp}), \tag{7.6}$$

where $q(\Omega, \vec{\wp}) \triangleq q_0(\Omega)\delta(\vec{\wp} - \vec{\wp}_0)$, $q_0(\Omega) = Q(\Omega)/dV$, $\vec{\wp}$ is an arbitrary location. This leads to a modified Helmholtz equation

$$\nabla^2 \mathrm{p} + k_\lambda^2 \frac{\partial^2 \mathrm{p}}{\partial t^2} = \jmath\Omega\rho_0 q(\Omega, \vec{\wp}). \tag{7.7}$$

The solution to (7.7) can be expressed with the eigenfunctions, i.e.

$$\mathrm{p}(\Omega, \vec{\wp}) = \sum_j D_j \mathrm{P}_j(\vec{\wp}), \tag{7.8}$$

where

$$D_j = \frac{1}{K_j} \iiint_V \mathrm{P}_j(\vec{\wp})\mathrm{p}(\Omega, \vec{\wp})dV, \tag{7.9}$$

and eigenfunctions depend on the room boundaries and are orthogonal, i.e.

$$\iiint_V \mathrm{P}_i(\vec{\wp})\mathrm{P}_j(\vec{\wp})dV = \begin{cases} K_j, & \text{for } i = j, \\ 0, & \text{for } i \neq j. \end{cases} \tag{7.10}$$

Expand also the source function with eigenfunctions,

$$q(\Omega, \vec{\wp}) = \sum_j C_j \mathrm{P}_j(\vec{\wp}), \tag{7.11}$$

where

$$C_j = \frac{1}{K_j} \iiint_V \mathrm{P}_j(\vec{\wp})q(\Omega, \vec{\wp})dV = \frac{1}{K_j} q_0(\Omega)\mathrm{P}_j(\vec{\wp}_0), \tag{7.12}$$

Inserting both (7.8) and (7.11) into (7.7), leads to

$$D_j = C_j \frac{\jmath\Omega}{k_\lambda^2 - k_j^2}. \tag{7.13}$$

Therefore, the sound pressure is

$$\mathrm{p}(\Omega, \vec{\wp}) = \jmath\Omega\rho_0 q_0(\Omega) \sum_j \frac{\mathrm{P}_j(\vec{\wp})\mathrm{P}_j(\vec{\wp}_0)}{K_j \cdot (k_\lambda^2 - |\vec{k}_j|^2)}, \tag{7.14}$$

where $\vec{k}_j$ is the j-th eigenvalue. The RTF in (2.10) is provided for a rectangular room with rigid boundaries [61].

B The Direct-path Subband Signal

From (3.2), the speech harmonic component is[1]

$$\begin{aligned} s_q^{(\hbar)}(t) =& A_q^{(\hbar)}(t) \cdot \cos\left(\hbar \cdot \omega_q \cdot t + \phi_q^{(\hbar)}(t)\right) \\ =& \frac{1}{2} A_q^{(\hbar)}(t) \cdot [e^{\jmath[\hbar \cdot \omega_q \cdot t + \phi_q^{(\hbar)}(t)]} + e^{-\jmath[\hbar \cdot \omega_q \cdot t + \phi_q^{(\hbar)}(t)]}] \end{aligned} \tag{7.15}$$

Using linear-phase filters, e.g. the gammatone filter, from (3.64),

$$\begin{aligned} g^{(b)}(t) &= \tilde{g}^{(b)}(t) \cdot \cos(2\pi f_c^{(b)} t) \\ &= \frac{1}{2} \cdot \tilde{g}^{(b)}(t) \cdot (e^{\jmath 2\pi f_c^{(b)} t} + e^{-\jmath 2\pi f_c^{(b)} t}) \end{aligned} \tag{7.16}$$

where

$$\tilde{g}^{(b)}(t) = (t + t_d)^{\vartheta - 1} e^{-2\pi f_b^{(b)}(t + t_d)} \tag{7.17}$$

From (3.2) and (7.16), when $\hbar \cdot \omega_q \approx 2\pi f_c^{(b)}$, the direct-path is given as follows:

$$\begin{aligned} & x_{d_i}^{(b)}(t) \approx [s_q^{(\hbar)}(t - t_{d_{qi}}) \cdot \mathrm{h}_{qi}(t_{d_{qi}})] * g^{(b)}(t) \\ & = \mathrm{h}_{qi}(t_{d_{qi}}) \cdot [\frac{1}{2} \cdot \tilde{g}^{(b)}(t) \cdot (e^{\jmath 2\pi f_c^{(b)} t} + e^{-\jmath 2\pi f_c^{(b)} t})] \\ & \quad * [\frac{1}{2} A_q^{(\hbar)}(t - t_{d_{qi}}) \end{aligned}$$

[1] Considering frequency domain meanings of convolution and complex exponentials for the Fourier transform and inverse Fourier transform.

$$
\begin{aligned}
&\cdot \left[e^{\jmath[\hbar\cdot\omega_q\cdot(t-t_{d_{qi}})+\phi_q^{(\hbar)}(t-t_{d_{qi}})]} + e^{-\jmath[\hbar\cdot\omega_q\cdot(t-t_{d_{qi}})+\phi_q^{(\hbar)}(t-t_{d_{qi}})]}\right] \\
&\approx \frac{1}{4}\mathrm{h}_{qi}(t_{d_{qi}})\cdot[A_q^{(\hbar)}(t-t_{d_{qi}}) * \tilde{g}^{(b)}(t)] \\
&\cdot \left[e^{\jmath[\hbar\cdot\omega_q\cdot(t-t_{d_{qi}})+\phi_q^{(\hbar)}(t-t_{d_{qi}})]} + e^{-\jmath[\hbar\cdot\omega_q\cdot(t-t_{d_{qi}})+\phi_q^{(\hbar)}(t-t_{d_{qi}})]}\right] \\
&= \frac{1}{2}\mathrm{h}_{qi}(t_{d_{qi}})\cdot[A_q^{(\hbar)}(t-t_{d_{qi}}) * \tilde{g}^{(b)}(t)]\cdot \\
&\quad \cos(\hbar\omega_q(t-t_{d_{qi}}) + \phi_q^{(\hbar)}(t-t_{d_{qi}})) \\
&= \tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t)),\ t \geq t_{qi}, \qquad (7.18)
\end{aligned}
$$

where

$$
\tilde{S}_{qi}^{(b)}(t) \triangleq \frac{1}{2}\cdot \mathrm{h}_{qi}(t_{d_{qi}})\cdot A_q^{(\hbar)}(t-t_{d_{qi}}) * \tilde{g}^{(b)}(t), \qquad (7.19)
$$

and

$$
\tilde{\phi}_{qi}^{(b)}(t) = \hbar\omega_q(t-t_{d_{qi}}) + \phi_q^{(\hbar)}(t-t_{d_{qi}}). \qquad (7.20)
$$

Particularly, in the case that the speech harmonic component is narrowband, and the center frequency is within the pass-band of the filter, (7.19) can be approximated as

$$
\tilde{S}_{qi}^{(b)}(t) \approx \frac{1}{2}\cdot \mathrm{h}_{qi}(t_{d_{qi}})\cdot A_q^{(\hbar)}(t-t_{d_{qi}})\cdot \tilde{G}^{(b)}(0), \qquad (7.21)
$$

where $\tilde{G}^{(b)}(f)$ is the Fourier transform of $\tilde{g}^{(b)}(t)$.

Thus from (7.18) and (7.21), the subband direct-path signal is an amplitude modulated sinusoid with slow-changing phase.

C The Expected Reflection Upper Bound

From (3.11),

$$\begin{aligned}
&\mathbb{E}[\lfloor x_{R_i}^{(b)}(t)\rfloor] = \mathbb{E}\big[\frac{1}{2}[x_{R_i}^{(b)}(t) + |x_{R_i}^{(b)}(t)|]\big] \\
=&\frac{1}{2}\mathbb{E}[|x_{R_i}^{(b)}(t)|] = \frac{1}{2}\mathbb{E}[|\mathrm{h}_R(t) * x_{d_i}^{(b)}(t)|] \\
\leq&\frac{1}{2}\mathbb{E}[|\mathrm{h}_R(t)| * |x_{d_i}^{(b)}(t)|] \\
=&\frac{1}{2}\mathbb{E}[|\mathrm{h}_R(t)|]*|x_{d_i}^{(b)}(t)| \\
=&\frac{1}{2}\mathbb{E}[|\mathrm{h}_R(t)|]*\big[\lfloor x_{d_i}^{(b)}(t)\rfloor + \lfloor -x_{d_i}^{(b)}(t)\rfloor\big],
\end{aligned} \tag{7.22}$$

where $\lfloor\cdot\rfloor$ keeps the non-negative part of signals while clipping negative signals to zero, i.e. $\lfloor x\rfloor = \frac{1}{2}(x+|x|), \forall x \in \mathbb{R}$. In the third line of (7.22) the equality holds when $\mathrm{h}_R(\tau)\cdot x_{d_i}^{(b)}(t-\tau)$ all have the same sign, $\forall \tau \geq \tau_{qi}$.

Thus from (7.18) and (7.22),

$$\begin{aligned}
&\mathbb{E}[\lfloor x_{R_i}^{(b)}(t)\rfloor] \\
\leq&\frac{1}{2}\mathbb{E}[|\mathrm{h}_R(t)|] * \big[\lfloor \tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t))\rfloor + \lfloor -\tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t))\rfloor\big] \\
=&\frac{1}{2}\mathbb{E}[|\mathrm{h}_R(t)|] * \big[\tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t))\big]_{t\in T_+^{(b)}} + \big[-\tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t))\big]_{t\in T_-^{(b)}}\big] \\
=&\frac{1}{2}\mathbb{E}[|\mathrm{h}_R(t)|] * [2\tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t))]_{t\in T_+^{(b)}} \\
=&\mathbb{E}[|\mathrm{h}_R(t)|] * \lfloor x_{d_i}^{(b)}(t)\rfloor,
\end{aligned} \tag{7.23}$$

where

$$\begin{aligned}
&[-\tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t))]_{t\in T_-^{(b)}} \\
\approx&[\tilde{S}_{qi}^{(b)}(t+\frac{T_b}{2})\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t+\frac{T_b}{2}))]_{t\in T_-^{(b)}} \\
=&[\tilde{S}_{qi}^{(b)}(t)\cdot\cos(\tilde{\phi}_{qi}^{(b)}(t))]_{t\in T_+^{(b)}} \\
=&\lfloor x_{d_i}^{(b)}(t)\rfloor,
\end{aligned} \tag{7.24}$$

$T_+^{(b)}$ is the set of time for non-negative $\cos(\tilde{\phi}_{qi}^{(b)}(t))$, while $T_-^{(b)}$ is the set of time for negative $\cos(\tilde{\phi}_{qi}^{(b)}(t))$. T_b is the short-term period of $\cos(\tilde{\phi}_{qi}^{(b)}(t))$.

D Recursive Averages of A Half-wave Rectified Periodic Signal

The recursive averages of the half-wave rectified sinusoid signal is calculated, the limit and an upper bound for λ less than but close to 1. Fig. 7.1 gives an intuition.

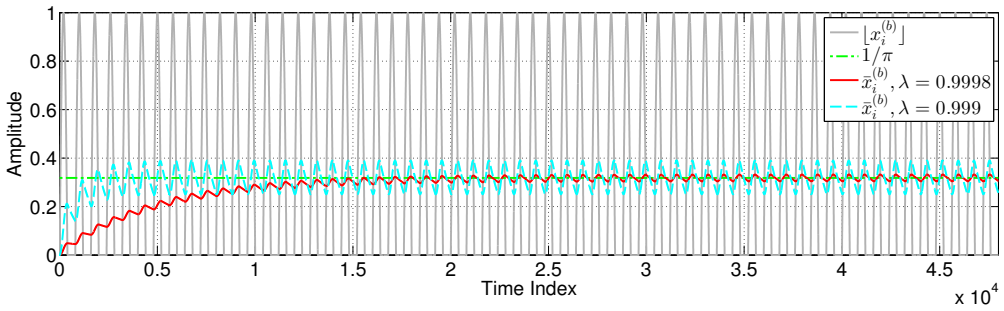

Figure 7.1: Intuition for the recursive average upper bound. Amplitude is fixed to 1 for the half-wave rectified sinusoid. ©2018 IEEE.

Assume a periodic signal $\lfloor x_i^{(b)}(l/f_s) \rfloor$ with period of integer $K_b \triangleq \text{round}(T_b \cdot f_s) > 0$, beginning at time $m_0 \in \mathbb{Z}$. From (3.18), the limit of the recursive average $\bar{x}_i^{(b)}[m]$ is:

$$\begin{aligned}
\lim_{m\to\infty} \bar{x}_i^{(b)}[m] &= \lim_{m\to\infty} (1-\lambda) \sum_{l=m_0}^{k} \lambda^{m-l} \cdot \lfloor x_i^{(b)}(l/f_s) \rfloor \\
&= \lim_{m\to\infty} \frac{1-\lambda}{K_b} \sum_{l=0}^{m} \lambda^{-l\cdot K_b} \sum_{l=m_0}^{m_0+K_b-1} \lambda^{m_0+K_b-1-l} \lfloor x_i^{(b)}(l/f_s) \rfloor \qquad (7.25) \\
&= \frac{1-\lambda}{1-\lambda^{K_b}} \cdot \frac{1}{K_b} \sum_{l=m_0}^{m_0+K_b-1} \lambda^{m_0+K_b-1-l} \lfloor x_i^{(b)}(l/f_s) \rfloor .
\end{aligned}$$

Using the cosine signal expression as in (7.18) assuming that $\tilde{S}_{qi}^{(b)}(t)$ is stable, the sum in (7.25) can be approximated using the integral, for λ close

to but less than 1, as in (3.23).

$$\begin{aligned}
&\int_0^{T_b} \lambda^{T_b - t} \cdot \tilde{S}_{qi}^{(b)}(t) \cdot \lfloor \cos(\tilde{\phi}_{qi}^{(b)}(t)) \rfloor dt \cdot f_s \\
&\approx \int_0^{T_b} \tilde{S}_{qi}^{(b)}(t) \cdot \lfloor \cos(\tilde{\phi}_{qi}^{(b)}(t)) \rfloor dt \cdot f_s \\
&= \tilde{S}_{qi}^{(b)}(t) \cdot \frac{T_b \cdot f_s}{\pi}.
\end{aligned} \tag{7.26}$$

Thus from (7.25) and (7.26) the limit approaches:

$$\lim_{m \to \infty} \bar{x}_i^{(b)}(m) \approx \frac{1-\lambda}{1-\lambda^{K_b}} \cdot \tilde{S}_{qi}^{(b)}(m/f_s) \cdot \frac{1}{\pi} \leq \tilde{S}_{qi}^{(b)}(m/f_s) \frac{1}{\pi}. \tag{7.27}$$

Thus $\tilde{S}_{qi}^{(b)}(m/f_s)/\pi$ is an upper bound of the recursive averages of the signal $\lfloor x_i^{(b)}(m/f_s) \rfloor$, which is also true when $\tilde{S}_{qi}^{(b)}(m/f_s)$ increases over time (e.g. speech onsets).

E The Localization of MCCC

Fig. 7.2 shows the localization test results from the MCCC method in the same test scenario as in Fig. 3.4 of Chapter 3. The source DOAs are 170° and 190°. It can be seen that the MCCC method cannot resolve the two close DOAs and the two sources are fused into one. There are considerable spurious peaks, and in most cases there is only about 0.1dB difference between the desired peaks and the spurious peaks. Thus the further test results from MCCC are not included in the detailed comparisons in Chapter 3.

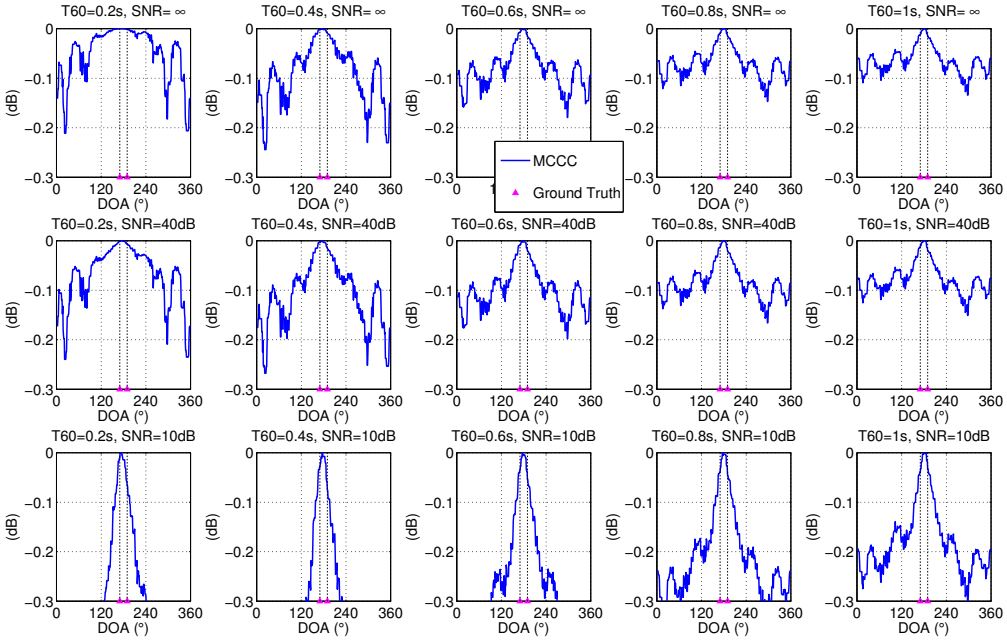

Figure 7.2: MCCC Localization Test.

F The Localization of SRP-PHAT

The localization results of the SPR-PHAT method for two closely located speakers in the same test scenario as in Fig. 3.4 of Chapter 3 are provided in Fig. 7.3. The source DOAs are 170° and 190°.

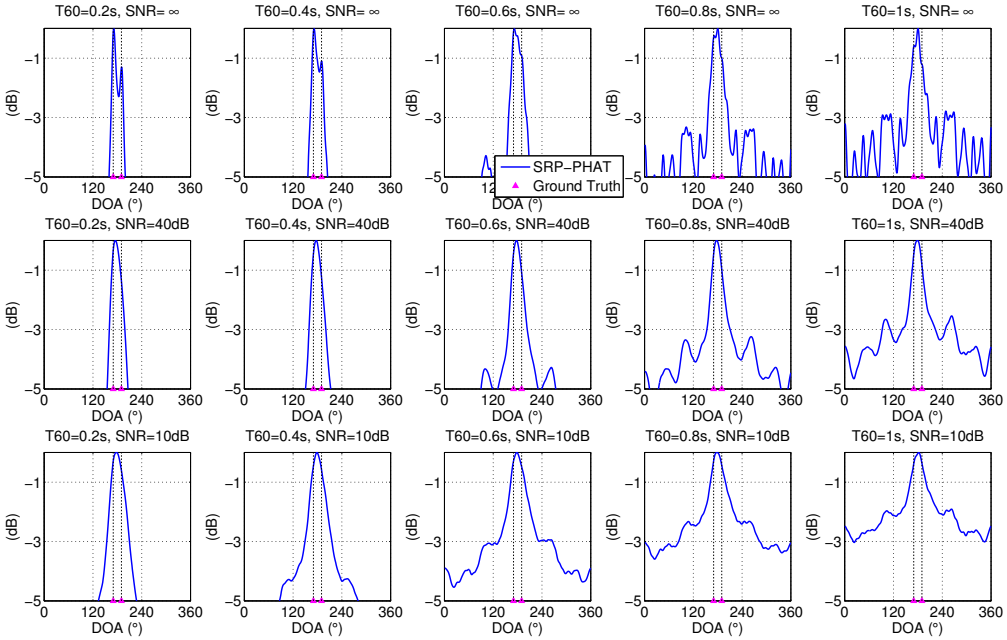

Figure 7.3: SRP-PHAT Localization Test.

It can be seen that in this test scenario:

- Although the differences between the main and the spurious peaks are only a few dBs, the SRP-PHAT is overall robust.
- The SRP-PHAT in general cannot differentiate the two closely located sources (except for the low reverberation and no noise cases), and the two sources are fused into one.
- There are more spurious peaks as the reverberation increases. Peaks are smeared as the noise increases. These observations follow the discussions in Section 3.5.

G The Localization of MUSIC

Fig. 7.4 shows the localization results of the MUSIC localization method in the same test scenario as in Fig. 3.4 of Chapter 3. The source DOAs are 170° and 190°. Although MUSIC is a well-accepted high resolution localization method, it can be seen that the resolution degrades as the reverberation or noise level increases. At high reverberation (e.g. $T_{60} > 0.4$s) the two sources are fused into one. Moreover, the peaks get wider as the reverberation or noise increases.

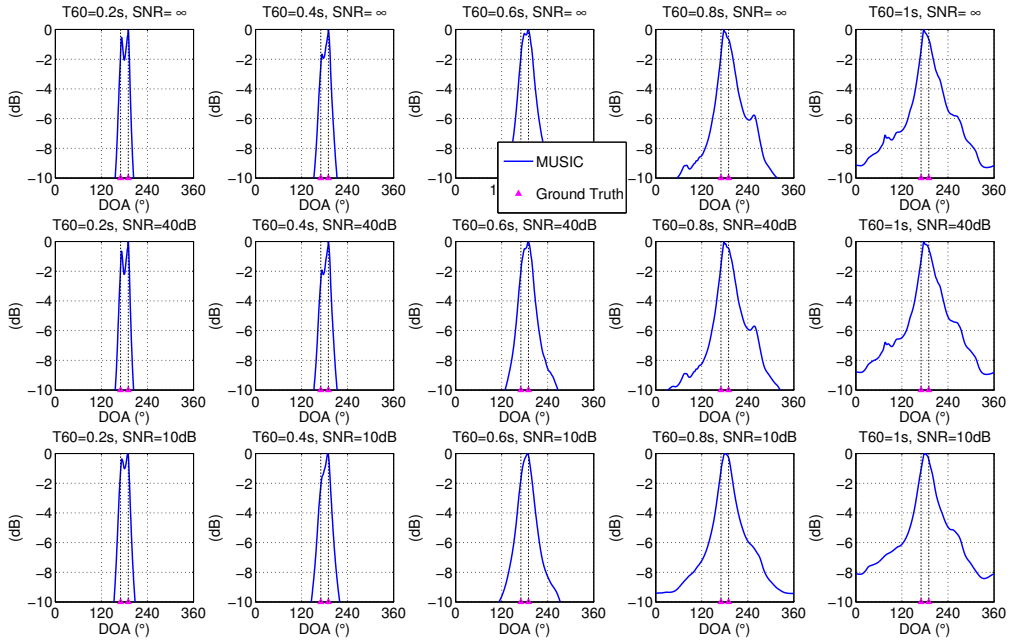

Figure 7.4: MUSIC Localization Test.

H The Localization of Onset-GSRP

Fig. 7.5 shows the localization results of the Onset-GSRP algorithm in the same test scenario as in Fig. 3.4 of Chapter 3. The source DOAs are 170° and 190°. The Onset-GSRP method cannot resolve the two closely located sources and fuses them into one. However, compared with the MCCC method in Fig. 7.2, the Onset-GSRP has improved performance, i.e. the main localization peak is distinct, and the spurious peaks much weaker than the main peak. It is also interesting to note that the peaks get wider as the reverberation or noise increases.

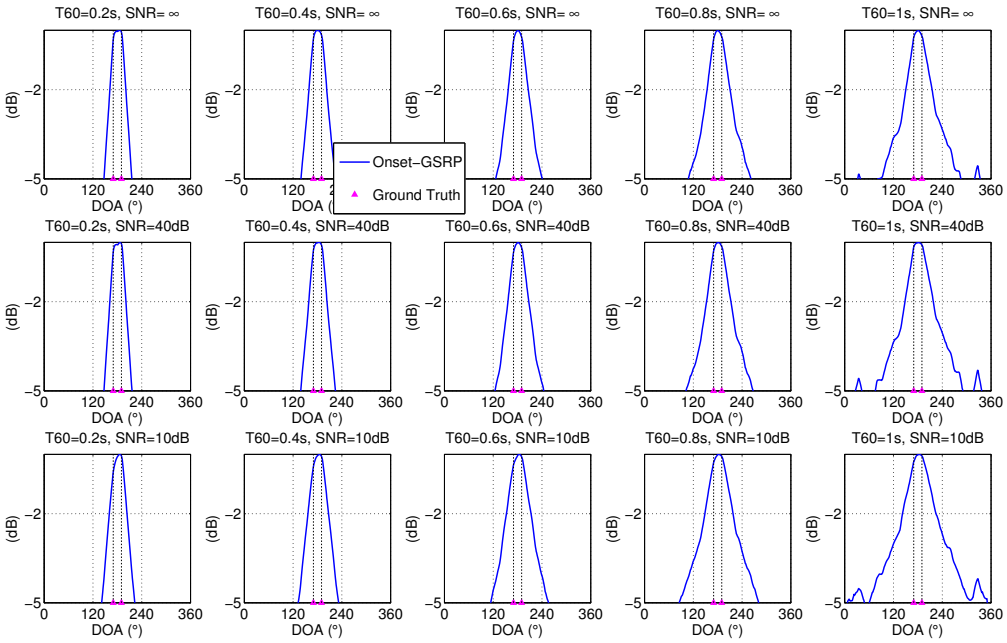

Figure 7.5: Onset-GSRP Localization Test.

I The Localization of Onset-MCC

Fig. 7.6 shows the localization results of the Onset-MCC method in the same test scenario as in Fig. 3.4 of Chapter 3. The source DOAs are 170° and 190°. The Onset-MCC method produces reliable peaks corresponding to ground truth speaker DOAs, except the spurious peaks at $T_{60} \geq 0.8$s and SNR=∞. Moreover, in comparison with Fig. 7.2, Fig. 7.3 and Fig. 7.4, the Onset-MCC has better resolution than the MCCC, SRP-PHAT and MUSIC.

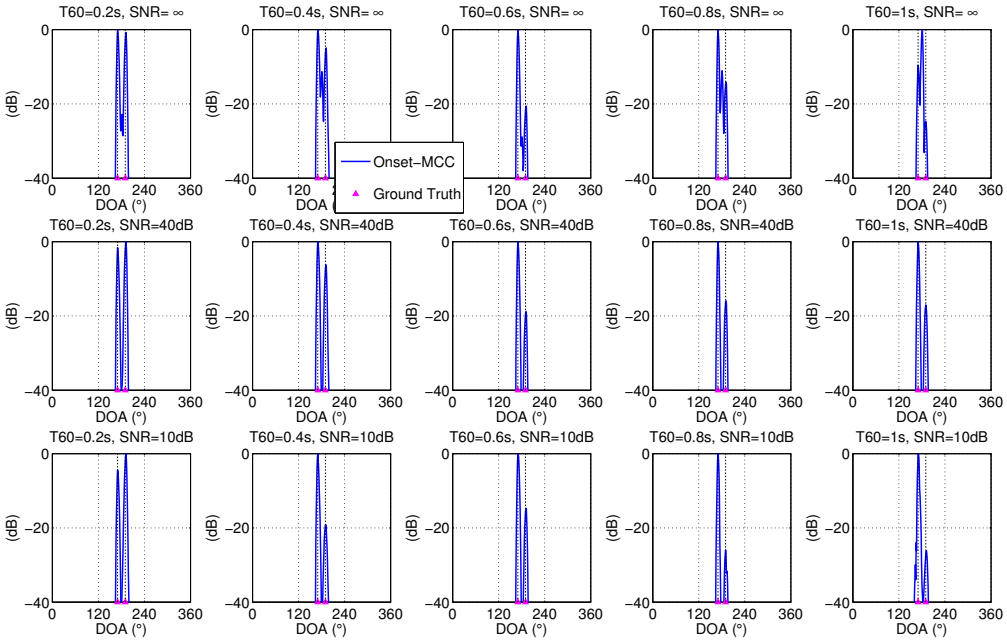

Figure 7.6: Onset-MCC Localization Test.

J The Localization of MCC-PHAT

Fig. 7.7 shows the localization results of the MCC-PHAT method in the same test scenario as in Fig. 3.4 of Chapter 3. The source DOAs are 170° and 190°. As shown in Fig. 7.7, removing spatially aliased pairs as in (3.59) does not significantly degrade the localization performance of the MCC-PHAT, compared to using all microphone pairs. The MCC-PHAT produces reliable peaks corresponding to ground truth speaker DOAs in most cases. Moreover, in comparison with Fig. 7.2, Fig. 7.3 and Fig. 7.4, the MCC-PHAT has better resolution than the MCCC, SRP-PHAT and MUSIC.

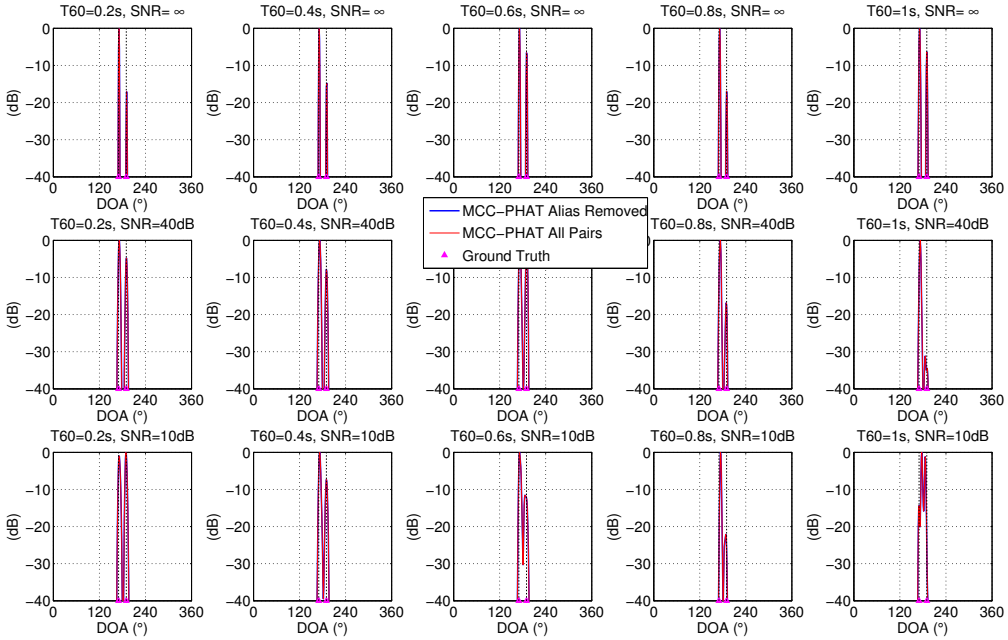

Figure 7.7: MCC-PHAT Localization Test.

K Computational Complexity of the Localization Algorithms

This section provides an analysis of the computational complexities of the proposed and existing baseline localization algorithms.

The test is carried out on a Thinkstation® with Xeon® E5-2640 CPU at 2.5GHz clock rate and 16GB DDR RAM, by running scripts using Matlab 2018b on the Windows 7 Professional operating system. The computational time is recorded when analyzing a 8-channel microphone recording of 4-second duration and 48kHz sampling rate. The result is shown in Fig. 7.8.

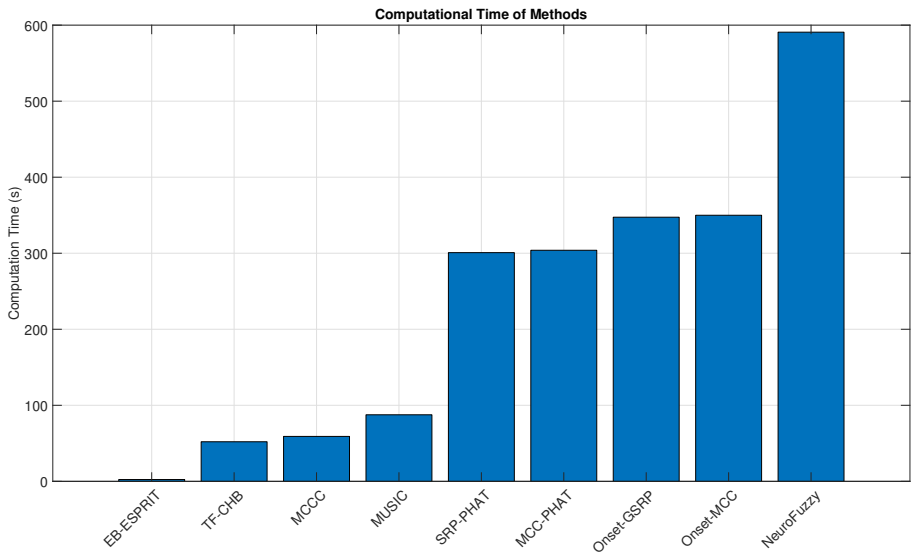

Figure 7.8: Computational Complexity Test of Localization Methods.

It is apparent that the EB-ESPRIT algorithm is by far the most efficient as it takes only 2.38 seconds in this case, while the Neuro-Fuzzy method takes the longest time of almost 600 seconds to complete. The TF-CHB and the MCCC algorithms have similar computational complexity (using about 50 seconds), while the MUSIC takes slightly longer (using 88 seconds). The SRP-PHAT, MCC-PHAT, Onset-GSRP and the Onset-MCC seem to require similar computational time of around 300 seconds. Note also that these last four

methods are reverberation-robust, where the MCC-PHAT and Onset-MCC methods can provide higher spatial resolution than the MUSIC method.

Fig. 7.9 provides the computational time for tracking location estimates from the real data as described in Chapter 4. The SMC-CPHD takes less than 2 seconds, which is much faster than the SMC-GLMB, but the latter provides identity estimates (labels) associated with the speaker locations. It takes longer for the SMC-GLMB to process the estimates from the MCC-PHAT method than the Onset-MCC, since the MCC-PHAT gives less miss-detections as discussed in Section 4.4.2.

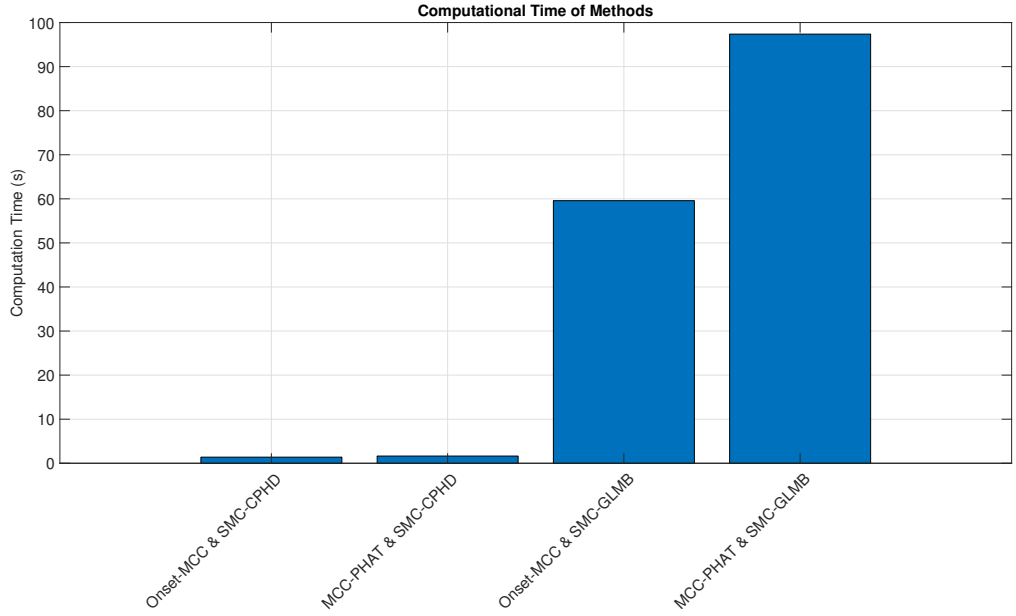

Figure 7.9: Computational Complexity Test of Tracking Methods.

L Permissions for Using © IEEE Papers

Title: Jointly Tracking and Separating Speech Sources Using Multiple Features and the Generalized Labeled Multi-Bernoulli Framework

Conference Proceedings: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Author: Shoufeng Lin

Publisher: IEEE

Date: April 2018

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

Bibliography

[1] Jaco Vermaak and Andrew Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01)*. IEEE, 2001, vol. 5, pp. 3021–3024.

[2] Darren B Ward, Eric Lehmann, Robert C Williamson, et al., "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[3] Ilyas Potamitis, Huimin Chen, and George Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[4] Wing-Kin Ma, Ba-Ngu Vo, Sumeetpal S Singh, and Adrian Baddeley, "Tracking an unknown time-varying number of speakers using tdoa measurements: a random finite set approach," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

[5] Nicoleta Roman and DeLiang Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[6] Shoufeng Lin, "Jointly tracking and separating speech sources using multiple features and the generalized labeled multi-bernoulli framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2018. ICASSP 2018.*

[7] L McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 305–317, 2005.

[8] Axel Plinge and Gernot A Fink, "Multi-speaker tracking using multiple distributed microphone arrays," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 614–618.

[9] Futoshi Asano, Masataka Goto, Katunobu Itou, and Hideki Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[10] Shoufeng Lin, Sven Erik Nordholm, Hai Huyen Dam, and Pei Chee Yong, "An adaptive low-complexity coherence-based beamformer," in *2013 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2013, pp. 263–266.

[11] Andrew H Jazwinski, *Stochastic Processes and Filtering Theory*, vol. 64, Academic Press, 1970.

[12] Simon J Julier and Jeffrey K Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[13] Yaakov Bar-Shalom, *Tracking and data association*, Academic Press Professional, Inc., 1987.

[14] Samuel S Blackman, "Multiple-target tracking with radar applications," *Dedham, MA, Artech House, Inc., 1986, 463 p.*, 1986.

[15] Ronald PS Mahler, "Multitarget bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, 2003.

[16] Ba-Ngu Vo, Sumeetpal Singh, and Arnaud Doucet, "Sequential monte carlo methods for multitarget filtering with random finite sets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1224–1245, 2005.

[17] Ronald Mahler, "Phd filters of higher order in target number," *IEEE Transaction on Aerospace and Electronic Systems*, vol. 43, no. 4, pp. 1523–1543, 2007.

[18] Ronald PS Mahler, *Statistical multisource-multitarget information fusion*, Artech House, Inc., 2007.

[19] Ba-Tuong Vo, Ba-Ngu Vo, and Antonio Cantoni, "Analytic implementations of the cardinalized probability hypothesis density filter," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3553–3567, 2007.

[20] Ba-Tuong Vo, Ba-Ngu Vo, and Antonio Cantoni, "The cardinality balanced multi-target multi-bernoulli filter and its implementations," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 409–423, 2009.

[21] Ba-Tuong Vo and Ba-Ngu Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3460–3475, 2013.

[22] Ba-Ngu Vo, Ba-Tuong Vo, and Dinh Phung, "Labeled random finite sets and the bayes multi-target tracking filter," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6554–6567, 2014.

[23] Stephan Reuter, Ba-Tuong Vo, Ba-Ngu Vo, and Klaus Dietmayer, "The labeled multi-bernoulli filter," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3246–3260, 2014.

[24] Irwin R Goodman, Ronald P Mahler, and Hung T Nguyen, *Mathematics of data fusion*, vol. 37, Springer Science & Business Media, 2013.

[25] Shoufeng Lin, Ba Tuong Vo, and Sven E Nordholm, "Measurement driven birth model for the generalized labeled multi-bernoulli filter," in *2016 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2016, pp. 94–99.

[26] Verne H MacDonald and Peter M Schultheiss, "Optimum passive bearing estimation in a spatially incoherent noise environment," *The Journal of the Acoustical Society of America*, vol. 46, no. 1A, pp. 37–43, 1969.

[27] Ralph O Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[28] Richard Roy and Thomas Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[29] BJORN Ottersten and Thomas Kailath, "Direction-of-arrival estimation for wide-band signals using the esprit algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 317–327, 1990.

[30] TL Tung, K Yao, D Chen, RE Hudson, and CW Reed, "Source localization and spatial filtering using wideband music and maximum power beamforming for multimedia applications," in *1999 IEEE Workshop on Signal Processing Systems, 1999. SiPS 99.* IEEE, 1999, pp. 625–634.

[31] Wei Liu and Stephan Weiss, *Wideband beamforming: concepts and techniques*, vol. 17, John Wiley & Sons, 2010.

[32] Elisabet Tiana-Roig, Finn Jacobsen, and Efrén Fernández Grande, "Beamforming with a circular microphone array for localization of environmental noise sources," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3535–3542, 2010.

[33] Abhaya Parthy, Nicolas Epain, André van Schaik, and Craig T Jin, "Comparison of the measured and theoretical performance of a broadband circular microphone array," *The Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 3827–3837, 2011.

[34] Ana M Torres, Maximo Cobos, Basilio Pueo, and Jose J Lopez, "Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1511–1520, 2012.

[35] Charles Knapp and Glifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[36] Jacob Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[37] Leslie S Smith and Steve Collins, "Determining itds using two microphones on a flat panel during onset intervals with a biologically inspired spike-based technique," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 15, no. 8, pp. 2278–2286, 2007.

[38] Soo-Yeon Lee and Hyung-Min Park, "Multiple reverberant sound localization based on rigorous zero-crossing-based itd selection," *IEEE Signal Processing Letters*, vol. 17, no. 7, pp. 671–674, 2010.

[39] Axel Plinge, Marius H Hennecke, and Gernot A Fink, "Robust neuro-fuzzy speaker localization using a circular microphone array," in *Proc. Int. Workshop on Acoustic Echo and Noise Control, Tel Aviv, Israel*. Citeseer, 2010.

[40] Heinz Teutsch and Walter Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2724–2736, 2006.

[41] Jacob Benesty, Jingdong Chen, and Yiteng Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 509–519, 2004.

[42] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone array signal processing*, vol. 1, Springer Science & Business Media, 2008.

[43] Joseph Hector DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, 2000.

[44] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180. Springer, 2001.

[45] Shoufeng Lin, "Reverberation-robust localization of speakers using distinct speech onsets and multichannel cross correlations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2098–2111, 2018.

[46] DeLiang Wang and Guy J Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE Press, 2006.

[47] William A Gardner, *Introduction to random processes*, vol. 71, McGraw-Hill, 1990.

[48] John R Deller Jr, John G Proakis, and John H Hansen, *Discrete time processing of speech signals*, Prentice Hall PTR, 1993.

[49] Athanasios Papoulis and S Unnikrishna Pillai, *Probability, random variables, and stochastic processes*, Tata McGraw-Hill Education, 2002.

[50] Ozgur Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[51] Lawrence R Rabiner, Ronald W Schafer, et al., "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[52] Aaron E Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[53] James V Candy, *Model-based signal processing*, vol. 36, John Wiley & Sons, 2005.

[54] Albert S Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

[55] Brian CJ Moore and Brian R Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[56] Julius O Smith III and Jonathan S Abel, "Bark and erb bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[57] RD Patterson, Ian Nimmo-Smith, John Holdsworth, and Peter Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, 1987, vol. 2.

[58] John Holdsworth, Ian Nimmo-Smith, Roy Patterson, and Peter Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[59] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[60] Lawrence E Kinsler, Austin R Frey, Alan B Coppens, and James V Sanders, "Fundamentals of acoustics," *Fundamentals of Acoustics, 4th Edition, by Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, James V. Sanders, pp. 560. ISBN 0-471-84789-5. Wiley-VCH, December 1999.*, p. 560, 1999.

[61] Heinrich Kuttruff, *Room acoustics*, Crc Press, 2014.

[62] Eric A Lehmann and Anders M Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[63] Manfred R Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal of the Audio Engineering Society*, vol. 35, no. 5, pp. 299–306, 1987. (Originally published in Acustica, vol. 4, pp. 594-600. (1954).).

[64] Manfred R Schroeder, "The "schroeder frequency"revisited," *The Journal of the Acoustical Society of America*, vol. 99, no. 5, pp. 3240–3241, 1996.

[65] Jean-Marc Jot, Laurent Cerveau, and Olivier Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Audio Engineering Society Convention 103*. Audio Engineering Society, 1997.

[66] Carl F Eyring, "Reverberation time in "dead" rooms," *The Journal of the Acoustical Society of America*, vol. 1, no. 2A, pp. 217–241, 1930.

[67] Manfred R Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 38, no. 2, pp. 359–361, 1965.

[68] Heiner Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary, "An improved algorithm for blind reverberation time estimation," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010, pp. 1–4.

[69] Durand R Begault and Leonard J Trejo, "3-d sound for virtual reality and multimedia," 2000.

[70] Eric A Lehmann and Anders M Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2010.

[71] R Schafer and L Rabiner, "Design and simulation of a speech analysis-synthesis system based on short-time fourier analysis," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 3, pp. 165–174, 1973.

[72] Jonathan Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[73] Jont B Allen and Lawrence R Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[74] Petre Stoica, Randolph L Moses, et al., "Spectral analysis of signals," 2005.

[75] Barry D Van Veen and Kevin M Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[76] Simon Doclo and Marc Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Transactions on Signal Processing*, vol. 51, no. 10, pp. 2511–2526, 2003.

[77] Y Kamp and J Thiran, "Chebyshev approximation for two-dimensional nonrecursive digital filters," *IEEE Transactions on Circuits and Systems*, vol. 22, no. 3, pp. 208–218, 1975.

[78] Sven Nordholm, Ingvar Claesson, and Per Eriksson, "The broad-band wiener solution for griffiths-jim beamformers," *IEEE Transactions on signal processing*, vol. 40, no. 2, pp. 474–478, 1992.

[79] Sven Nordebo, Ingvar Claesson, and Sven Nordholm, "Weighted chebyshev approximation for the design of broadband beamformers using quadratic programming," *IEEE Signal Processing Letters*, vol. 1, no. 7, pp. 103–105, 1994.

[80] SE Nordholm, Volke Rehbock, KL Tee, and Sven Nordebo, "Chebyshev optimization for the design of broadband beamformers in the near field," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 1, pp. 141–143, 1998.

[81] Otis Lamont Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[82] Lloyd Griffiths and CW Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[83] Sven E Nordholm, Hai H Dam, Chiong C Lai, and Eric A Lehmann, "Broadband beamforming and optimization," in *Academic Press Library in Signal Processing*, vol. 3, pp. 553–598. Elsevier, 2014.

[84] Boaz Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 135–143, 2005.

[85] Or Nadiri and Boaz Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 10, pp. 1494–1505, 2014.

[86] Milton Abramowitz and Irene A Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, vol. 55, Courier Corporation, 1965.

[87] MA Doran, Eyal Doron, and Anthony J Weiss, "Coherent wide-band processing for arbitrary array geometry," *IEEE Transactions on Signal Processing*, vol. 41, no. 1, pp. 414, 1993.

[88] Alan W Rudge, K Milne, and A David Olver, *The handbook of antenna design*, vol. 16, IET, 1982.

[89] Earl G Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*, Elsevier, 1999.

[90] Hong Wang and Mostafa Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[91] Mati Wax and Thomas Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[92] Petre Stoica and Arye Nehorai, "Music, maximum likelihood, and cramer-rao bound," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 5, pp. 720–741, 1989.

[93] Harry L Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*, John Wiley & Sons, 2004.

[94] Petre Stoica and Arye Nehorai, "Music, maximum likelihood, and cramer-rao bound: further results and comparisons," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2140–2150, 1990.

[95] Russell Jeffers, Kristine L Bell, and Harry L Van Trees, "Broadband passive range estimation using music," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 3, pp. III–2921.

[96] Anders Johansson and Sven Nordholm, "Robust acoustic direction of arrival estimation using root-srp-phat, a realtime implementation," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. IEEE, 2005, vol. 4, pp. iv–933.

[97] Heinz Teutsch, "Wavefield decomposition using microphone arrays and its application to acoustic scene analysis," 2005.

[98] Heinz Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, vol. 348, Springer, 2007.

[99] Haohai Sun, Heinz Teutsch, Edwin Mabande, and Walter Kellermann, "Robust localization of multiple sources in reverberant environments using eb-esprit with spherical microphone arrays," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 117–120.

[100] G Clifford Carter, "Coherence and time delay estimation," *Proceedings of the IEEE*, vol. 75, no. 2, pp. 236–255, 1987.

[101] Julius S Bendat and Allan G Piersol, "Random data analysis and measurement procedures," 2000.

[102] Jingdong Chen, Jacob Benesty, and Yiteng Arden Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 026503, 2006.

[103] Axel Plinge, Marius H Hennecke, and Gernot A Fink, "Reverberation-robust online multi-speaker tracking by using a microphone array and casa processing," in *International Workshop on Acoustic Signal Enhancement; Proceedings of IWAENC 2012;*. VDE, 2012, pp. 1–4.

[104] Mingyang Wu, DeLiang Wang, and Guy J Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[105] Anssi Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255–266, 2008.

[106] Byung Suk Lee and Daniel P.W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Proc. INTERSPEECH, Portland, OR, USA*, Sep, 2012.

[107] YC Ho and RCKA Lee, "A bayesian approach to problems in stochastic estimation and control," *IEEE Transactions on Automatic Control*, vol. 9, no. 4, pp. 333–339, 1964.

[108] Rudolph Emil Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[109] Andrew H Jazwinski, *Stochastic processes and filtering theory*, Courier Corporation, 2007.

[110] Neil J Gordon, David J Salmond, and Adrian FM Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," in *IEE Proceedings F (Radar and Signal Processing)*. IET, 1993, vol. 140, pp. 107–113.

[111] Arnaud Doucet, "On sequential simulation-based methods for bayesian filtering," 1998.

[112] Samuel S Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 1, pp. 5–18, 2004.

[113] Roy L Streit and Tod E Luginbuhl, "Probabilistic multi-hypothesis tracking," Tech. Rep., NAVAL UNDERWATER SYSTEMS CENTER NEWPORT RI, 1995.

[114] Songhwai Oh, Stuart Russell, and Shankar Sastry, "Markov chain monte carlo data association for general multiple-target tracking

problems," in *Decision and Control, 2004. CDC. 43rd IEEE Conference on*. IEEE, 2004, vol. 1, pp. 735–742.

[115] Ba Tuong Vo, *Random finite sets in multi-object filtering*, Ph.D. thesis, Citeseer, 2008.

[116] Steven M Kay, "Fundamentals of statistical signal processing, volume i: estimation theory," 1993.

[117] Crispin W Gardiner et al., *Handbook of stochastic methods*, vol. 3, Springer Berlin, 1985.

[118] Hung Gia Hoang, Ba Tuong Vo, and Ba-Ngu Vo, "A fast implementation of the generalized labeled multi-bernoulli filter with joint prediction and update," in *Information Fusion (Fusion), 2015 18th International Conference on*. IEEE, 2015, pp. 999–1006.

[119] Ba-Ngu Vo, Ba-Tuong Vo, and Hung Gia Hoang, "An efficient implementation of the generalized labeled multi-bernoulli filter," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 1975–1987, 2017.

[120] B Ristic, D Clark, Ba-Ngu Vo, and Ba-Tuong Vo, "Adaptive target birth intensity for phd and cphd filters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 2, pp. 1656–1668, 2012.

[121] Stephan Reuter, Daniel Meissner, Benjamin Wilking, and Klaus Dietmayer, "Cardinality balanced multi-target multi-bernoulli filtering using adaptive birth distributions," in *Information Fusion (FUSION), 2013 16th International Conference on*. IEEE, 2013, pp. 1608–1615.

[122] Dominic Schuhmacher, Ba-Tuong Vo, and Ba-Ngu Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008.

[123] Bhaskar D Rao and KV Sl Hari, "Performance analysis of root-music," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1939–1949, 1989.

[124] Marius Pesavento, Alex B Gershman, and Martin Haardt, "Unitary root-music with a real-valued eigendecomposition: A theoretical and experimental performance study," *IEEE transactions on signal processing*, vol. 48, no. 5, pp. 1306–1314, 2000.

[125] Jacek P Dmochowski, Jacob Benesty, and Sofiene Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, 2007.

[126] Brad Rakerd and William M Hartmann, "Localization of sound in rooms, iii: Onset and duration effects," *The Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1695–1706, 1986.

[127] Jie Huang, Noboru Ohnishi, and Noboru Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Transactions on Instrumentation and Measurement*, vol. 46, no. 4, pp. 842–846, 1997.

[128] Simon Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects*. Citeseer, 2006, vol. 120, pp. 133–137.

[129] Marco Kuhne, Roberto Togneri, and Sven Nordholm, "Robust source localization in reverberant environments based on weighted fuzzy clustering," *IEEE signal processing letters*, vol. 16, no. 2, pp. 85–85, 2009.

[130] Nguyen Thi Ngoc Tho, Shengkui Zhao, and Douglas L Jones, "Robust doa estimation of multiple speech sources," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2287–2291.

[131] Martin Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[132] Ruth Y Litovsky, H Steven Colburn, William A Yost, and Sandra J Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[133] Ray Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *The Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702–711, 1986.

[134] Petros Maragos, James F Kaiser, and Thomas F Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE transactions on signal processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[135] Jean-Dominique Polack, "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2-4, pp. 235–244, 1993.

[136] Patrick A Naylor and Nikolay D Gaubitch, *Speech dereverberation*, Springer Science & Business Media, 2010.

[137] Richard Lyon, "A computational model of binaural localization and separation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.* IEEE, 1983, vol. 8, pp. 1148–1151.

[138] Ray Meddis and Lowel O'Mard, "A unitary model of pitch perception," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–1820, 1997.

[139] Tero Tolonen and Matti Karjalainen, "A computationally efficient multipitch analysis model," *IEEE transactions on speech and audio processing*, vol. 8, no. 6, pp. 708–716, 2000.

[140] Eric A Lehmann, Anders M Johansson, and Sven Nordholm, "Reverberation-time prediction method for room impulse responses simulated with the image-source model," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 159–162.

[141] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic data consortium*, vol. 10, no. 5, pp. 0, 1993.

[142] Fotios Talantzis, "An acoustic source localization and tracking framework using particle filtering and information theory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1806–1817, 2010.

[143] Ba-Ngu Vo and Ba-Tuong Vo, "An implementation of the multi-sensor generalized labeled multi-bernoulli filter via gibbs sampling," in *Information Fusion (Fusion), 2017 20th International Conference on*. IEEE, 2017, pp. 1–8.

[144] Ba-Ngu Vo, Sumeetpal Singh, and Wing Kin Ma, "Tracking multiple speakers using random sets," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04)*. IEEE, 2004, vol. 2, pp. ii–357.

[145] Hiroshi Sawada, Ryo Mukai, Shoko Araki, and Shoji Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[146] Taesu Kim, Hagai T Attias, Soo-Young Lee, and Te-Won Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 1, pp. 70–79, 2007.

[147] Vaninirappuputhenpurayil Gopalan Reju, Soo Ngee Koh, and Yann Soon, "Underdetermined convolutive blind source separation via time–frequency masking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 101–116, 2010.

[148] Douglas A Reynolds, "An overview of automatic speaker recognition technology," in *2002 IEEE international conference on Acoustics, speech, and signal processing (ICASSP)*. IEEE, 2002, vol. 4, pp. IV–4072.

[149] David Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[150] Xuejing Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, vol. 1, pp. I–333.

[151] Sira Gonzalez and Mike Brookes, "Pefac-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[152] Dongmei Wang, Chengzhu Yu, and John HL Hansen, "Robust harmonic features for classification-based pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 952–964, 2017.

[153] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[154] J Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 998–1003, 1982.

[155] J Ianniello, Ehud Weinstein, and Anthony Weiss, "Comparison of the ziv-zakai lower bound on time delay estimation with correlator performance," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'83*. IEEE, 1983, vol. 8, pp. 875–878.

[156] Petr Tichavsky, Carlos H Muravchik, and Arye Nehorai, "Posterior cramér-rao bounds for discrete-time nonlinear filtering," *IEEE Transactions on signal processing*, vol. 46, no. 5, pp. 1386–1396, 1998.

[157] Tony Gustafsson, Bhaskar D Rao, and Mohan Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[158] Branko Ristic, Sanjeev Arulampalam, and Neil James Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*, Artech house, 2004.

[159] Irwin Pollack and James M Pickett, "Cocktail party effect," *The Journal of the Acoustical Society of America*, vol. 29, no. 11, pp. 1262–1262, 1957.

[160] Simon Haykin and Zhe Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[161] Adelbert W Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.